Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Can Current Archival Data Constrain the Age Evolution of the Sub-Neptune Fraction Around Cool Stars? An Inconclusive Injection-Calibrated Test

Zubayer Hasan Shaad

zubayerhasanshaad@gmail.com

Government Tolaram College https://orcid.org/0009-0002-2322-8553

Md Mahdin Al Abrar

Chandpur Government College https://orcid.org/0009-0009-9645-152X

Research Article

Keywords: Exoplanets, Exoplanet evolution, M dwarf stars, Astrostatistics, Exoplanet catalogs

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11835824/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Can Current Archival Data Constrain the Age Evolution of the Sub-Neptune Fraction Around Cool Stars? An Inconclusive Injection-Calibrated Test

Zubayer Hasan Shaad (ORCID: 0009-0002-2322-8553)
Md Mahdin Al Abrar (ORCID: 0009-0009-9645-152X)

Government Tolaram College, Narayanganj, Bangladesh

Email: zubayerhasanshaad@gmail.com

Chandpur Government College, Chandpur, Bangladesh

Email: mdmahdinalabrar@gmail.com

Abstract

Current archival data cannot settle whether the radius valley evolves with stellar age around cool stars, and the obstacle is statistical power rather than a demonstrated absence of evolution. We ask whether a literature-sized age effect can be discriminated from no effect with a calibrated gate that injects null and signal realizations into the same noise and selection structure as the data, then requires the real fit to separate the two arms. The analysis is defined along three axes that we state explicitly: it targets a heterogeneous late-K/early-M sample with $T_{\rm eff} < 4200$ K rather than the M3+ population, it adjudicates quantitatively between a physical effect and a power-limited measurement, and it enforces an explicit PASS/FAIL injection decision rather than a nominal significance threshold. The completeness-weighted FGK control ($N = 2{,}135$ planets across 1,588 hosts, ESS $= 1{,}645$, 1,000 replicates per arm) fails the gate: the real coefficient (-0.0296 ± 0.1061 per IQR) lies at the 45.6th percentile of the null, with a likelihood ratio of 0.68 and a signed arm separation of -1.01 null standard deviations. For the heterogeneous cool sample ($N = 417$, ESS $= 140$) the signed separation is only -0.19, so the two arms are effectively the same distribution and the verdict is also FAIL. A power grid detects the injected -0.14 per SD effect in only 9.4% of realizations at the current scale, rising to 23.2% [19.6–27.2] at $8\times$ (~3,337 planets) and 60.6% [56.2–64.9] for a twofold stronger effect; with the empirical null quantile the corresponding rates are 7.2% at $1\times$, and 19.8% and 55.8% at $8\times$. The kinematic age proxy is validated for FGK hosts ($\rho = 0.50$, $n = 79$); for cool hosts the pipeline gyro sample gives only $\rho = 0.25$ ($n = 12$), while an external TIME

Table crossmatch validates ordinal age ordering ($\rho = 0.37, p = 0.0023, n = 65$) without calibrating an absolute zero point. A model-dependent extrapolation places the conventional 80% power scale between 1.7×10^4 and 4.7×10^4 planets. More independent hosts, per-host ages, and per-star completeness products are required, not a larger planet count alone.

Keywords: Exoplanets, Exoplanet evolution, M dwarf stars, Astrostatistics, Exoplanet catalogs

1 Introduction

The radius valley, the bimodal gap near 1.5–2.0 $R_\oplus$ in the distribution of close-in small planets, is one of the most robust demographic features in the exoplanet census (Fulton et al. 2017; Van Eylen et al. 2018). Its physical origin is debated: leading explanations include atmospheric mass loss driven by XUV photoevaporation (Owen and Wu 2017; Lopez and Fortney 2013) or core-powered mass loss (Ginzburg et al. 2018; Gupta and Schlichting 2019), both of which predict that the relative populations on either side of the valley should evolve with time as envelopes are gradually stripped, and formation epoch mechanisms such as gas-poor formation (Lee and Chiang 2016; Lopez and Rice 2018), which predict that the valley is imprinted at birth and does not evolve.

These predictions are testable: if atmospheric stripping dominates, older planetary systems should show a depleted sub-Neptune population relative to younger ones, because marginal sub-Neptunes lose their envelopes over Gyr timescales and migrate across the valley into the super-Earth peak.

1.1 What Is and Is Not Established for FGK Hosts

The observational picture among FGK hosts is more mixed than is sometimes stated, and the distinction matters for how a control sample should be interpreted.

The evidence is strongest for the *number ratio* channel. Berger et al. (2020) used isochrone ages for Kepler hosts and found that the sub-Neptune to super-Earth ratio decreases with age. David et al. (2021) reported a consistent trend using gyrochronological ages, and Chen et al. (2022) extended this with a refined age analysis. Sandoval et al. (2021) found the super-Earth to sub-Neptune detection ratio to rise over roughly 1 to 10 Gyr in the California-Kepler Survey at about the two sigma level, reporting an increase from 0.82 to 0.97 across a 3 Gyr split. Kamulali et al. (2026), in a homogeneous reanalysis based on SWEET-Cat spectroscopic parameters, report a super-Earth to sub-Neptune number ratio rising from $0.51^{+0.11}_{-0.08}$ below 3 Gyr to $0.64^{+0.11}_{-0.11}$ above it.[1]

[1] The published Kamulali et al. (2026) analysis reports 1221 stars and 1405 planets in its abstract but 779 stars and 893 planets in its final valley summary, an internal sample-count discrepancy that we flag rather than silently reconcile. We adopt their published

The evidence is considerably weaker for the *valley location* and *sub-Neptune size* channels. Kamulali et al. (2026) measure an age slope of $0.00^{+0.05}_{-0.03}$ for the valley location in their two-dimensional fit and $0.07^{+0.03}_{-0.04}$ in their four-dimensional fit, and conclude explicitly that age trends remain inconclusive given the uncertainties in stellar age estimates, while Ho and Van Eylen (2023) report an age slope compatible with zero.

More recent work has sharpened this disagreement. A homogeneous analysis of 769 planets around 558 FGK hosts (Chen et al. 2026) examines valley morphology against age, activity, chemistry, and birth radius jointly, finds the valley weak below roughly 3 Gyr and evolving non-monotonically, and attributes the intermediate-age weakening to thin and thick disc mixing; because those covariates vary together, age is not an independent control in that analysis. A selection-conditioned, two-component mixture analysis of 746 planets around 510 FGK hosts with gyrochronological ages (Greco et al. 2026) instead finds the intrinsic super-Earth mixture weight rising from 0.43 to 0.67, with 90 percent of the increase complete before 1.9 Gyr ($t_{90} < 1.9$ Gyr); superposing planets across orbital periods dilutes the valley contrast from its intrinsic depth of 0.39 to 0.28, so conditioning on period recovers the deeper intrinsic structure rather than erasing it. By contrast, the California-Kepler Survey X found no significant variation in sub-Neptune size with stellar age over 1 to 10 Gyr and no significant size-metallicity correlation (Petigura et al. 2022), and an inverse-detection-efficiency study of Kepler FGK hosts found no significant occurrence-age trend (Sayeed et al. 2025). The observational baseline is therefore best described as a rising number ratio within the first couple of gigayears followed by a plateau, with no robust monotonic metallicity direction. Chen and colleagues attribute their intermediate-age dip to population mixing rather than a true reversal, so a monotonic injection remains an adequate bracket for the underlying envelope-loss trend: it captures the early rise and the plateau, and a population-mixing dip would only reduce the effective contrast further. Our power analysis should therefore be read as an optimistic bracket, not as a model of the detailed non-monotonic morphology.

We therefore do not treat FGK age evolution as an established effect against which a pipeline can be calibrated by reproduction. We treat it as a literature motivated effect size whose magnitude is itself uncertain, and we carry that uncertainty explicitly in Section 3.4. Our model targets the mixing fraction and is therefore aimed at the number ratio channel, which is the channel in which the literature evidence is strongest.

1.2 The M Dwarf Question

Whether this age evolution extends to M dwarfs is a separate and currently open question. M dwarfs are cooler, smaller, and less luminous than FGK stars, with different XUV luminosity histories and convective structures. Theoretical predictions for mass loss driven valley evolution around M dwarfs are model-dependent: some photoevaporation

---

number ratios as given. The 136 young and 687 old counts used in our effect-size derivation (Section 3.4) come from their Figure 7 and are a separate, age-split accounting.

models predict a weaker or delayed signal (Owen and Wu 2017), while core-powered mass loss models make different predictions for the low-mass stellar regime (Gupta and Schlichting 2020). The observational picture is sparse. Gaidos et al. (2024) found the sub-Neptune to super-Earth ratio to fall from 2.38 to 0.58 across a median split near 3.86 Gyr among 3200–4200 K hosts, applying Kepler DR25 detection-efficiency corrections and validating recovery with synthetic injections. They did not, however, apply a null versus signal decision gate to that contrast. It is also not settled that a radius based classification is the right instrument in this regime: Luque and Pallé (2022) argue that planets around M dwarfs separate by bulk density into rocky and water rich populations rather than by radius into a valley, and Cloutier and Menou (2020) show that the M dwarf valley's character depends strongly on orbital period.

The age information available for cool hosts is itself limited. Population-synthesis work predicts that the valley fades toward lower stellar mass as water-rich planets become common (Venturini et al. 2024), while TESS-based analyses find a shallower valley around low-mass hosts (Ho et al. 2024). In the warmer part of the cool-star regime a valley near 1.6 to 1.7 Earth radii is still measured (Parashivamurthy and Mulders 2025), so the relevant boundary appears to lie near 0.3 to 0.4 solar masses rather than at the effective-temperature cut adopted here.

The question was sharpened considerably by Gillis et al. (2026), who surveyed 8,134 mid-to-late M dwarfs, roughly 0.1–0.4 $M_{\odot}$, observed by TESS with injection recovery completeness corrections and recovered a unimodal radius distribution peaking at $1.25 \pm 0.05 R_{\oplus}$, in contrast to the bimodal structure around earlier type hosts. They proposed a formation based explanation, in which the valley around hotter hosts is sculpted primarily by atmospheric evolution while M dwarf planets form in a regime where the valley never develops, without testing the age axis. If that formation only hypothesis is correct, there should be no age evolution of the relevant M dwarf demographics; if atmospheric stripping operates but more weakly, an attenuated age signal could remain. The literature provides no quantitative test distinguishing these possibilities.

We note at the outset that our M dwarf selection is $T_{\rm eff} < 4200$ K and is therefore dominated by late K and early M hosts. It is not the approximately M3 and later population in which Gillis et al. (2026) locate the valley disappearance, so the title question is scoped to this heterogeneous < 4200 K sample, and we return to this limitation in Section 5.1.

1.3 This Paper

This paper addresses a specific question: can current archival samples discriminate a literature-sized age effect on the radius valley around cool stars from a null effect? We approach this not as a detection paper but as a statistical adjudication. The study is deliberately scoped along three axes, which we state firmly here: a cool-host sample with $T_{\rm eff} < 4200$ K; a quantitative adjudication of whether a physical effect or limited statistical power governs the result; and an explicit PASS/FAIL injection gate. We develop that gate by injecting known null and literature-sized synthetic signals into the same noise structure

and requiring the real data to discriminate between them. We validate the age proxy directly, test explicitly for selection effects, apply the gate to both an FGK control and the cool sample, and quantify how sample size and age information affect the statistical path to resolution. We do not claim the first mixture study of this kind, the first age test around M dwarfs, the first diagnosis of period mixing, or a resolution of the underlying formation mechanism.

2 Data

2.1 Sample Construction

We construct our planet samples from the NASA Exoplanet Archive using the Planetary Systems (`ps`) table, querying via TAP with the following selection criteria: `default_flag` $= 1$, `pl_controv_flag` $= 0$, `discoverymethod` $=$ `'Transit'`, planet radius $0.5 < R_p < 4.0$ $R_\oplus$, and orbital period $P < 100$ d. These cuts isolate the small planet population spanning the radius valley while excluding controversial dispositions and restricting to the period range where transit surveys have meaningful sensitivity.

We emphasize that this is a catalog of *confirmed* planets with heterogeneous provenance, vetting histories, and stellar parameter sources. It is not a uniform detection sample of the kind required for a full occurrence rate calculation, which would need the complete searched stellar sample together with per-star detection efficiency, vetting completeness, and reliability. All fractions reported here are therefore conditional on confirmation, and the completeness weighting described in Section 3.1 corrects for relative detectability within the confirmed catalog rather than recovering an occurrence rate.

We divide the sample by host star effective temperature into two subsamples: a cool sample, $T_{\rm eff} < 4200$ K, yielding 419 planets around 303 host stars; and a broader FGK parent sample, $4200 \leq T_{\rm eff} \leq 7000$ K, yielding 2,880 planets around 2,150 host stars. The 2,150 figure refers to this full parent sample, not to the fitted control. The fitted gate uses the inner join with Gaia kinematics and completeness weights: 417 planets around 301 hosts for the cool sample (the two planets around Kepler-1410 and Kepler-1652 have no usable Gaia astrometry) and 2,848 planets around 2,123 hosts for the FGK parent (32 planets around 27 hosts lost on the join); the adopted FGK control further restricts to Kepler-prefixed hosts, giving 2,135 planets around 1,588 hosts with ESS $= 1{,}645$ after the 95th-percentile weight trim.

The cool-star cut spans a heterogeneous range, and its mass composition matters for how the result should be read. The 303 hosts have a median effective temperature of 3683 K, with quartiles of 3415 and 3867 K; 23.4% of hosts lie below 3400 K and 1.3% below 3000 K. Among the 282 hosts with a usable stellar mass, 35.8% lie below 0.4 $M_\odot$ and 20.6% below 0.3 $M_\odot$. Weighted by planet, 23.9% of the 419 planets orbit hosts below 3400 K and 36.1% of planets with a usable host mass orbit hosts below 0.4 $M_\odot$. The sample therefore straddles the contested 0.3–0.4 $M_\odot$ transition where the valley is reported to fade. We

adopt a 1.85 $R_\oplus$ classification for this sample (Table 2); the 1.6–1.7 $R_\oplus$ valley reported for warmer cool stars does not apply across this heterogeneous regime, and the 1.25 $R_\oplus$ value is the super-Earth mode measured by Gillis et al. (2026), not a valley location. We do not resolve the transition here; we report the distribution so the scope of the verdict is explicit (Figure 1).

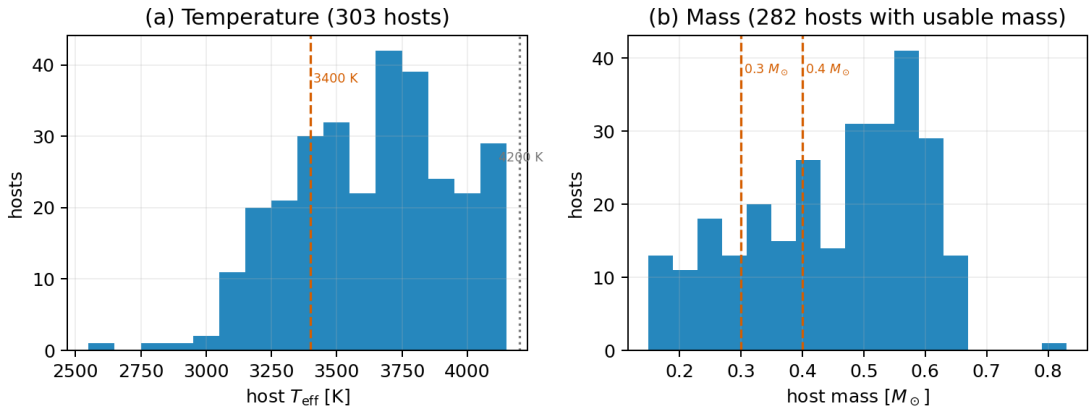

Figure 1. Host distribution of the cool sample. (a) Effective temperature for 303 hosts; dashed line at 3400 K marks the cooler regime where the valley is contested, and the dotted line at 4200 K is the sample cut. (b) Stellar mass for the 282 hosts with usable values; dashed lines at 0.3 and 0.4 $M_\odot$ bracket the transition where atmospheric-loss valley contrast fades. A substantial minority of hosts lies in the contested regime, so the verdict is scoped to a heterogeneous cool sample rather than to a valley-present population.

2.2 Gaia DR3 Crossmatch and Astrometric Quality

We crossmatch all host stars against Gaia DR3 (`gaiadr3.gaia_source`) using a cone search, then apply an offline quality filter requiring a positive parallax and adequate proper motion measurements. Of the 303 M dwarf hosts, 301 have usable Gaia astrometry; the offline quality gate rejects 2 spurious matches (the entries associated with Kepler-1410 and Kepler-1652). Of the 2,150 FGK hosts, 2,123 have usable astrometry after the same quality gate.

We note two limitations of this astrometric treatment. No cut on the Gaia renormalized unit weight error (RUWE) was applied in production. RUWE > 1.4 flags solutions poorly described by a single-star model, most often unresolved binaries, which are precisely the systems whose proper motions and therefore $v_{\rm tan}$ values are least reliable. A post-hoc sensitivity gate re-queried Gaia DR3 for 2,424 host source IDs (301 cool plus 2,123 FGK) and reran the analysis under RUWE ≤ 1.4 and parallax S/N > 5: 35/417 M dwarf planets (23/301 hosts) and 178/2135 FGK planets (143/1588 hosts) are dropped, arm separations move from -0.19 to -0.25 (M) and -1.01 to -0.84 (FGK), and both verdicts remain FAIL (Appendix C). Applying the same cuts before fitting gives a real pooled $\beta = -0.00280 \pm 0.36017$ for the cool sample and -0.04868 ± 0.11720 for the FGK control, both consistent

with zero. Distances are taken as inverse parallaxes without a parallax zero point correction, which is adequate at the median distances of these samples but would matter for a more distant host set. Across the 2,424 quality-table hosts the fractional inverse-parallax distance error has a median of 0.0088 and a 95th percentile of 0.0462; for the hosts with parallax S/N ≤ 5 the median is 0.35, and these are precisely the hosts the production cut removes. No parallax-uncertainty propagation was applied; distances for low-S/N parallaxes should be treated with an appropriate prior in future work.

The archived acquisition script used sequential 1.5 arcsec cone searches against Gaia DR3 and selected the nearest source by angular separation. Proper motions were used as reported by Gaia DR3, with no propagation to a common reference epoch. These choices were not recorded explicitly in the original acquisition manifest, so we document them here.

2.3 Kinematic Age Proxies

For each host with usable Gaia astrometry, we compute tangential velocity ($v_{\tan}$) from parallax and proper motion and, where radial velocities are available, full heliocentric Galactic velocities (U, V, W). Radial velocities are available for 237 of the 301 M dwarf hosts and 1,429 of the 2,123 FGK hosts. Tangential velocity serves as the primary kinematic age proxy: kinematically hotter (higher $v_{\tan}$) populations are statistically older (Aumer and Binney 2009), but $v_{\tan}$ is a population level age proxy rather than a precise age measurement for an individual star.

2.4 Validation of the Age Proxy

Because the inferential claims of this paper depend on how much age information $v_{\tan}$ actually carries, we validate it directly against an independent age estimator. For the 79 FGK hosts that have both a usable gyrochronological age (Section 2.5) and usable astrometry, the Spearman correlation between $v_{\tan}$ and gyrochronological age is

$$\rho = +0.50, \quad p = 3.0 \times 10^{-6}, \quad n = 79,$$

recomputed from the frozen host tables and archived with the analysis outputs,[2] and $\rho = +0.471$ ($p = 7.6 \times 10^{-5}$) if the 14 hosts censored at the gyrochronology grid boundary are excluded. This is a substantially stronger validation than the small number of literature young anchors available previously, and it is a direct calibration of the proxy used here.

Two consequences follow. First, attenuation: a regression coefficient defined per standard deviation of true age appears in $v_{\tan}$ units at approximately ρ times its value, so an effect defined against a spectroscopic or isochrone age scale is roughly halved when expressed against this proxy. The injected effect size adopted in Section 3.4 should be read with this

[2] The numeric record is `proxy_validation_fgk_recomputed.json` in the deposited release.

in mind: -0.14 per SD of $v_{\tan}$ is an upper bound, since the same physical effect on a true-age scale would appear at roughly half that value where the proxy is validated.

Second, and more awkwardly, the same test applied to M dwarfs gives $\rho = +0.25$ ($p = 0.43$) on only 12 hosts, of which 4 of 12, or 33.3%, are censored at the grid boundary. *The age proxy is therefore validated for the control sample and lacks an absolute calibration for the sample of interest within our pipeline's gyrochronology domain*. This is consistent with independent work showing that on the lower main sequence effective temperature and surface gravity do not discriminate age well enough to assemble coeval ensembles, so kinematic ages there degrade to population-level constraints without an additional age indicator (Lu and Pinsonneault 2026). We do not transport the FGK attenuation factor across, and we treat this as a first-order limitation of the M dwarf analysis rather than a technicality.

The independent TIME Table gyrochronological ages give a much larger validation set. Across the 65 cool hosts with finite published ages (Section 2.5), the rank correlation with $v_{\tan}$ is $\rho = +0.372$ ($p = 0.0023$, $n = 65$); the 62 hosts additional to our pipeline gyro set give $\rho = +0.382$ ($p = 0.0022$). The association is therefore present and significant in the regime where our own gyro ages were unavailable, even though the zero point of the kinematic proxy remains uncalibrated there. We report this as a validation of age *ordering* only, not of an absolute age scale, and we do not feed these ages into the primary gate because they fall outside the Angus et al. (2019) validity domain.

We note that hosts with archive rotation periods are biased toward active and therefore younger stars, so the FGK correlation is range restricted and should be read as a lower bound on the true association.

2.5 Gyrochronological Ages

We compute rotation based ages using the `gyrointerp` package implementing the Angus et al. (2019) slow sequence gyrochronology model. The model's validity domain requires $3800 < T_{\rm eff} < 6200$ K and rotation period $P_{\rm rot} \leq 45$ d; below 3800 K the model returns NaN, a boundary discovered and enforced during the project's audit process. Archive rotation period uncertainties are used with a [5%, 35%] floor ceiling clip.

Of the 303 M dwarf hosts, 66 have archive rotation periods, but only 12 yield usable gyrochronological ages (median 2.50 Gyr), with 33.3% censored at the 3 Gyr model grid boundary. Of the 2,150 FGK hosts, 88 have archive rotation periods and 79 yield usable ages (median 1.93 Gyr), with 18% censored. The low M dwarf yield reflects the limited calibration of gyrochronology for cool stars, and it is the direct reason the validation in Section 2.4 is weak for M dwarfs. A literature crossmatch against the McQuillan et al. (2014) Kepler rotation catalog recovered one additional M dwarf rotation period not present in the archive (Kepler-1646, matched at 0.26 arcsec separation).

A further crossmatch against the TIME Table of Gaidos et al. (2023) matched 71 of our 303 cool hosts. Of these, 66 are additional to the 12 hosts with pipeline gyrochronological ages,

and 62 have finite published ages. The full machine-readable TIME Table contains 249 rows, comprising 227 age estimates, 8 upper limits, and 14 exclusions, archived at Zenodo DOI 10.5281/zenodo.7578269. These are literature gyrochronological ages derived from Gaidos's own temperature-dependent gyrochrones, calibrated over 3200–4200 K. They do not automatically satisfy the validity domain of the Angus et al. (2019) model used in our pipeline (3800–6200 K and $P_{\rm rot} \leq 45$ d), so we do not count the 62 as new pipeline-usable ages. Their stated caveats also apply: a median age error of 20% and a mode of 14%; a peak below 3 Gyr that the authors identify as a calibration artefact; a 75 K effective-temperature error that propagates to a 0–17% age error; an untested $n = 0.62$ spin-down exponent; and the possibility of serious age errors for binary hosts (Figure 2).

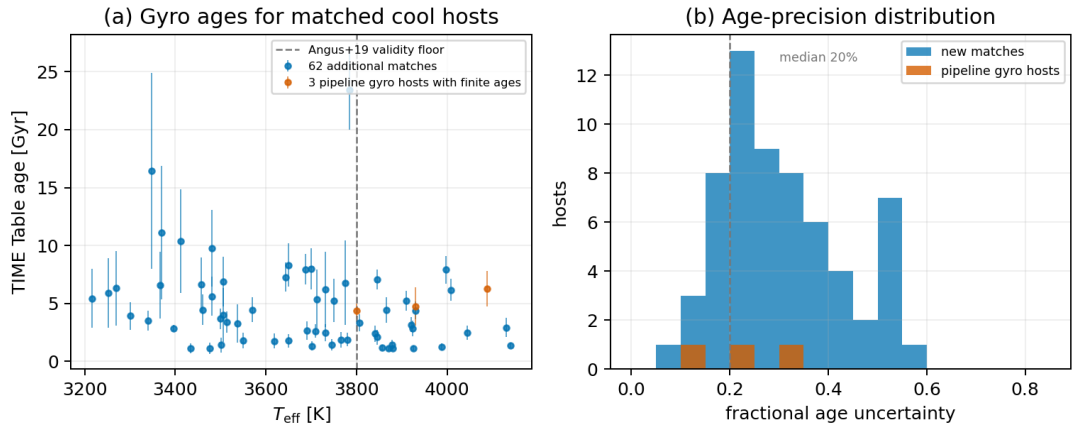

Figure 2. TIME Table crossmatch. (a) Published gyrochronological age against effective temperature for 65 matched hosts with finite published ages (62 additional to our gyro set plus 3 overlapping hosts with finite ages); orange points are those 3 overlapping hosts, blue points are the additional matches. Two further overlapping hosts are upper limits rather than finite ages. The dashed line at 3800 K marks the lower validity edge of the Angus et al. (2019) model used in our pipeline. (b) Fractional age uncertainty; the dashed line marks the 20 percent median quoted by Gaidos et al. (2023). The additional ages enlarge the potential rotation-aged sample but are not used in the primary gate because they fall outside our model's validity domain.

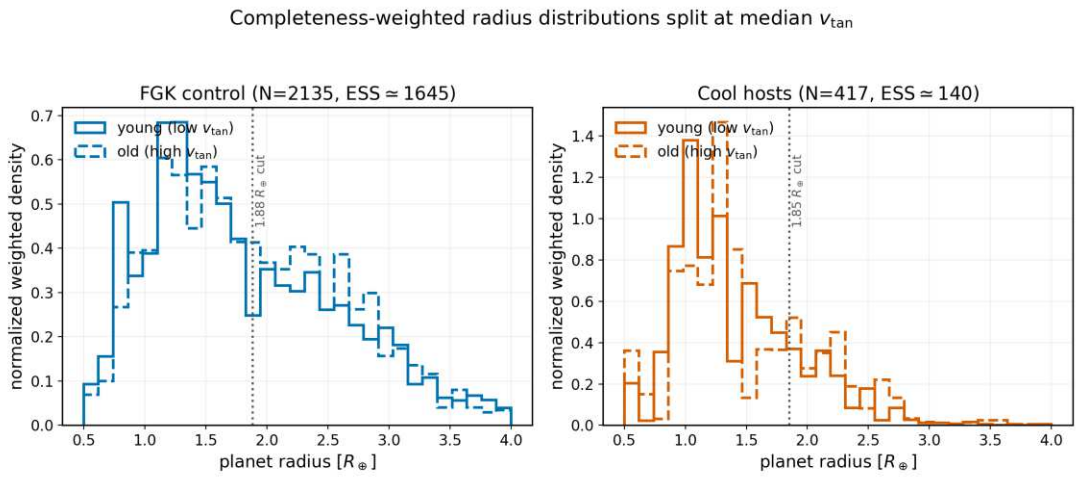

Figure 3. Completeness-weighted planet radius distributions split at the median tangential velocity, for the FGK control (left) and the cool sample (right). Dotted lines mark the classification thresholds used in the binned tests. The FGK panel shows the expected structure with a shoulder near the valley; the cool panel is far noisier, reflecting an effective sample size of 140 rather than the nominal 417 planets (Table 2).

3 Methods

The estimand throughout is the relative within-catalog slope of the sub-Neptune mixing fraction per interquartile range of $v_{\tan}$. It is not an absolute occurrence rate, and no claim about the true planet occurrence around these stars follows from it. Recovering an absolute rate would require the full searched stellar sample together with per-star detection efficiency, vetting completeness, and reliability, which this confirmed planet catalog does not provide.

3.1 Completeness Weighting

Transit detection probability varies strongly across the planet parameter space and between survey facilities. Any analysis comparing sub-Neptune fractions across stellar or planetary populations must account for this because uncorrected comparisons are biased toward easily detected (large, short period) planets and can manufacture spurious demographic trends.

For Kepler observed hosts, we use the real per-star 6 hour Combined Differential Photometric Precision (CDPP) values from the Kepler DR25 stellar table (`rrmscdpp06p0`), retrieved from the NASA Exoplanet Archive `keplerstellar` table and matched via KEPID. This provides empirically measured photometric noise for 1,691 of 1,695 Kepler prefixed hosts in the combined sample (99.8% coverage), spanning 2,161 of the 2,848 FGK parent planets observed by Kepler.

Gillis et al. (2026) performed injection recovery completeness corrections for their M dwarf TESS sample; however, no machine readable sensitivity table is linked from their published materials, and their sample covers only mid to late M dwarfs observed by TESS, with partial overlap to our host set. Their product was therefore not directly applicable.

For K2 and TESS hosts, where per-star CDPP products are not available in a single uniform table, we compute a first-order individual detection efficiency model (IDEM) from physical observables: transit depth from planet radius and stellar radius, transit duration from stellar density, number of transits from mission specific observing baselines, and signal to noise ratio evaluated against magnitude scaled CDPP with a logistic detection probability function centered at SNR ≈ 7.5, floored at 0.02 to prevent weight divergence.

The primary FGK control restricts to Kepler prefixed hosts only (2,135 planets), so the dominant completeness weight component derives from real DR25 CDPP rather than the cruder non Kepler proxy. We apply inverse detection probability weighting ($w = 1/P_{\mathrm{det}}$), normalize to mean unity, and trim at the 95th percentile of w. Trimming is a bias–variance trade rather than a pure gain: it discards the tail of the weight distribution that formally carries the correction for the least detectable planets, and we therefore report untrimmed and 90th percentile variants in Tables 1 and 2 so that the cost is visible.

The progression from the naive all-survey scheme (ESS 487, 17% retention, 2.42× SE inflation) through Kepler only untrimmed (ESS 511, 24%, 2.04×) to the adopted Kepler only trimmed scheme (ESS 1,645, 77%, 1.14×) removes the extreme high-weight tail: effective sample size rises by a factor of 3.4 and the SE inflation falls from 2.42× to 1.14×. Retention near 30% (the M sample retains 33%, and its untrimmed scheme 29%) leaves the estimator effectively uninformative, and an absolute occurrence claim would require the full selection function rather than inverse detection weights alone.

The 95th percentile is not the retention maximising choice: trimming more aggressively raises ESS further (88% at the 90th percentile, 97% at the 80th) because it removes more of the weight variance. It does so by discarding more of the correction, however, and the estimand moves accordingly. In the FGK control the completeness-weighted sub-Neptune fraction runs 0.488, 0.446, 0.420, 0.396 and 0.382 at the 80th, 90th, 95th and 99th percentile trims and untrimmed, while the fitted age coefficient runs -0.0157 ± 0.1017, -0.0296 ± 0.1061, -0.0537 ± 0.1239 and -0.0955 ± 0.2639 at the 90th, 95th and 99th percentile trims and untrimmed (no refit at the 80th trim). The coefficient moves monotonically away from zero as trimming is relaxed, while its standard error grows faster still, so every variant remains comfortably consistent with zero. We adopt the 95th percentile as a compromise between retained information and discarded correction, and note that the verdict of the calibrated gate is unchanged across this range even though the point estimate is not.

Table 1. Completeness weighting schemes for the FGK control sample.

Scheme	N	ESS	Retention	SE infl.	f_{SN}
Unweighted (reference)	2,135	2,135	100%	1.00×	0.532

Scheme	N	ESS	Retention	SE infl.	f_{SN}
Naive all survey (rejected)	2,848	487	17%	2.42×	0.437
Kepler only, untrimmed	2,135	511	24%	2.04×	0.382
Kepler only, trimmed 90th	2,135	1,868	88%	1.07×	0.446
Kepler only, trimmed (adopted)	2,135	1,645	77%	1.14×	0.420

ESS is the effective sample size under each weighting scheme; retention is ESS/N. SE inflation is the factor by which standard errors are inflated relative to an unweighted analysis of the same N. f_{SN} is the sub-Neptune fraction under a 1.88 $R_{\oplus}$ classification.

Table 2. Completeness weighting schemes for the cool sample.

Scheme	N	ESS	Retention	SE infl.	f_{SN}
Unweighted (reference)	417	417	100%	1.00×	0.408
Weighted, untrimmed	417	121	29%	1.85×	0.221
Weighted, trimmed 90th	417	169	41%	1.57×	0.260
Weighted, trimmed (adopted)	417	140	33%	1.73×	0.234
DR25 CDPP subset, trimmed	126	117	93%	1.04×	0.289

As Table 1, with f_{SN} under a 1.85 $R_{\oplus}$ classification. The cool sample retains 33% of its nominal information under the adopted scheme, against 77% for the FGK control, so it behaves as a 140 planet sample rather than a 417 planet one.

The contrast between the two tables is important and is not visible from planet counts alone. The FGK control retains 77% of its information; the cool sample retains 33%. Moreover, 70% of the cool-sample weights derive from the IDEM proxy (182 V band, 108 Kepler magnitude, 1 default) rather than from real DR25 CDPP (126), and the weighted sub-Neptune fraction of the cool sample moves from 0.408 unweighted to 0.234 weighted, against 0.532 to 0.420 for FGK. The weights are therefore doing substantially more work on the cool sample, with larger maximum influence on the result, and on a weaker empirical footing. Uncertainty in the IDEM itself, including its logistic detection threshold, its probability floor, and its magnitude-scaled CDPP assumption, is not propagated into the injection ensembles; the calibrated gate therefore conditions on the adopted weights rather than marginalising over their model uncertainty. This is a limitation of the present calibration that we do not resolve here.

Restricting the M sample to planets with real DR25 CDPP weights raises retention to 93% at a cost of only 16% of the effective sample size, but the resulting subset is not representative: its hosts are hotter (median T_{eff} 3799 K against 3550 K, $p = 1.9 \times 10^{-10}$) and kinematically cooler (median $v_{\tan}$ 27.9 against 34.0 km s^{-1}, $p = 2.8 \times 10^{-3}$), because it is the Kepler observed tail of the sample and therefore moves away from the mid to late M population of interest. We retain the full weighted sample as primary and report the DR25 subset as a sensitivity check in Section 4.4.

3.2 Hierarchical Mixture Model

We model the observed planet radius distribution as a two-component Gaussian mixture (super-Earths and sub-Neptunes), where the mixing fraction, the probability that a planet belongs to the sub-Neptune component, is the science target. The mixing fraction is parameterized as a logistic function of stellar age proxy ($v_{\tan}$, standardized) and covariates $\log_{10}(P)$ and [Fe/H] where available, with the age proxy coefficient β_{age} as the primary parameter of interest. The likelihood is evaluated via Gauss-Hermite quadrature over radius measurement uncertainty, with component means and widths constrained to data measured values to prevent degeneracy: component starts are the medians and standard deviations of radii below and above the valley floor, giving $(\mu_{\text{SE}}, \sigma_{\text{SE}}, \mu_{\text{SN}}, \sigma_{\text{SN}}) = (1.40, 0.31, 2.56, 0.49)$ for FGK (valley 1.88 $R_{\oplus}$) and $(1.28, 0.31, 2.29, 0.44)$ for M dwarfs (valley 1.85 $R_{\oplus}$), with fits restricted to ± 0.06 in means and ± 0.12 in log-widths.

This model improves on a Bernoulli classifier applied to a fixed radius threshold: it treats radius as a continuous observed outcome with measurement error, avoids information loss from dichotomization near the valley, and accommodates overlap between the super-Earth and sub-Neptune populations.

Constraining the component means and widths is a deliberate scope restriction as well as a numerical stability measure, and it has a consequence worth stating plainly: the model can express a change in the super-Earth to sub-Neptune *ratio* with age, but it cannot express a shift in the valley *location*. Our test therefore addresses the channel in which the literature evidence is strongest (Section 1.1) and is insensitive by construction to the channel in which it is contested.

Host-level clustering.

The 419 M dwarf planets orbit 303 hosts and the 2,880 FGK planets orbit 2,150 hosts; planets sharing a star are not independent draws because they share host-level properties including age, formation history, and disk composition. Point estimates are obtained under a working independence likelihood with the host random intercept scale fixed at zero, a deliberate choice made to avoid the documented weak identifiability competition between the age slope and a host random effect. Reported standard errors are host-level cluster robust: the sandwich covariance aggregates per planet scores to per-host scores before forming the meat matrix. The injection machinery preserves the same structure by resampling whole hosts and drawing one shared host effect per host ($\sigma_u = 1.5$ logits, the host-overdispersion scale of the v1 hierarchical fit, itself a $\sqrt{2}$-deflated estimate of ≈ 2.0–2.3; larger true heterogeneity would lower power further). A one-planet-per-host robustness check is reported in Section 5.6.

3.3 Calibrated Gate Methodology

The central methodological contribution of this work is a calibrated statistical gate designed to adjudicate whether an observed fit discriminates between competing

hypotheses. For a given sample and weighting scheme, three quantities are generated: (1) the real data fit, yielding β_{age}; (2) a null injected ensemble of synthetic realizations with $\beta_{\mathrm{inj}} = 0$; and (3) a signal injected ensemble with β_{inj} set to the literature motivated value derived in Section 3.4. Each synthetic realization bootstraps whole hosts from the real sample, carrying their real radius uncertainties and covariates, so the injected ensembles inherit the same noise and correlation structure as the data.

The gate criterion is:

- PASS: the real β_{age} is consistent with the signal injected distribution within its 95% envelope and statistically distinguishable from the null injected distribution.
- FAIL (inconclusive): one or both conditions are not met, so the data cannot discriminate between the two hypotheses.

Operationally, PASS requires the real coefficient to lie inside the signal arm's central 95% interval and in the lower tail of the null injected distribution. Because the hypothesised atmospheric-loss effect is negative, the distinguishability test against the null is one-sided and lower-tail: it requires $P(\text{null} \leq \text{real}) \leq 0.05$, that is, the real coefficient must fall in the lower 5% of the null injected ensemble (implemented in the archived calibration code described in the Data and Reproducibility Availability section). Signal containment uses the empirical 2.5th and 97.5th quantiles of the signal arm. Placement intervals are Wald intervals, $\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/n}$, clipped to the unit interval. The ± 1.6 percentage points quoted at 1,000 replicates per arm is one binomial standard error, not a 95% interval; the corresponding 95% interval would be ± 3.1 percentage points. Percentiles are therefore quoted to 0.1% precision only as rankings, not as sharp thresholds.

The emphasis on calibration over analytic significance testing follows recent methodological work on ensemble null tests (Bayer and Spergel 2026), the sample-to-sample variability of p values (Halsey et al. 2015), safeguards for discovery claims and the publication of null results (Junk and Lyons 2020), and the identifiability and consistency of two-component mixtures (Ndaoud 2022; Mallik 2026).

The criterion is intentionally strict. A FAIL verdict is not a null result; it is a statement that the sample lacks the power to adjudicate the question under the tested specification.

We additionally report two summary quantities that we find more legible than percentile placements alone. The first is the *likelihood ratio* between the two injected densities evaluated at the real coefficient, estimated by Gaussian kernel density (SciPy `gaussian_kde`, Scott's bandwidth) fitted separately to each arm's replicates; values near unity indicate that the data are uninformative between the hypotheses. The estimator is reproduced by the archived likelihood-ratio check: 0.68 for FGK pooled from the archived 1000/arm replicates and 1.74 for FGK stratified from the archived 400/arm replicates, with M pooled at 0.93 in production (1.03 from the 200/arm appendix ensemble). The second is the *arm separation*, the difference between the signal and null injected medians expressed in units of the null standard deviation, with the convention $(\text{signal} - \text{null})/\text{null SD}$ used

consistently in tables and prose; because the signal median is more negative, the signed values are negative. This summarizes how distinguishable the two hypotheses are under the tested model for a given sample. It is a diagnostic of available power, not the decision criterion (Section 4.4); the verdict rests on the real coefficient's placement in the null ensemble.

3.4 Derivation of the Injected Effect Size

The signal injected value is derived from Kamulali et al. (2026) as follows. Their super-Earth to sub-Neptune number ratio is 0.51 for hosts younger than 3 Gyr and 0.64 for older hosts. As sub-Neptune fractions these are

$$f_{\mathrm{SN}} = \frac{1}{1 + N_{\mathrm{SE}}/N_{\mathrm{SN}}} = 0.662 \ \text{ and } \ 0.610,$$

a logit difference of $0.447 - 0.674 = -0.227$ across the age split (old minus young). Dividing by the separation between the two age groups in standard deviations of the age variable gives the per standard deviation coefficient. For a median split of a normal distribution the two group means are separated by 1.60 SD, giving -0.14 per SD. Kamulali et al. (2026) in fact split 136 young (≤ 3 Gyr) from 687 old (> 3 Gyr) hosts, a total of 823, so the young hosts form 16.5% of the combined sample. For a normal age distribution the separation between the young and old group means at a split quantile of 0.165 is 1.80 SD, giving an implied slope of -0.126 per SD.[3] We retain $\beta_{\mathrm{inj}} = -0.14$ per SD from the median-split assumption as our primary injected value; the -0.126 alternative is about 10% smaller and leaves all FAIL verdicts unchanged given arm separations of -1.01 (FGK) and -0.19 (M) null SD.

Two caveats attach to this number. The uncertainties on 0.51 and 0.64 are ± 0.11 and heavily overlapping, so the effect size is itself poorly determined; a beta-distribution injection that propagates this uncertainty is reported in the appendix. And Kamulali et al. (2026) sample 4400–7100 K hosts, so transporting this calibration to cool hosts is an assumption rather than a measurement. We retain the number-ratio calibration rather than the selection-conditioned mixture weight of Greco et al. (2026) because our estimand is a within-catalog slope per IQR of $v_{\tan}$, not an intrinsic mixture weight; the two are related but not identical. The effect-size sensitivity span from Sandoval to Gaidos nevertheless brackets a Greco-sized change.

Results are reported per interquartile range of $v_{\tan}$. The empirical IQR/SD ratio is 1.236 for the FGK control and 1.263 for the cool sample; these differ from the Gaussian value of 1.35 because $v_{\tan}$ is a positive, right skewed quantity (skewness 1.59 and 0.96 respectively).

---

[3] The quantity 136/687 $= 0.198$ is a young-to-old count ratio, not a percentile. The relevant percentile is the young fraction of the total, 136/823 $= 16.5\%$, and we use the 16.5th percentile throughout.

3.5 Adversarial Audits and Analytical Corrections

The analysis pipeline underwent two independent adversarial audits aimed specifically at finding errors that could alter the inference, followed by a third review pass. The audits are part of the methodological record rather than a post hoc correction step, and all corrections described below were applied before the final results were reported.

The most significant issues identified were a mixture identifiability collapse, in which the original fixed radius Bernoulli formulation allowed variance to absorb the mixing signal and attenuated β_{age} by approximately 35%, corrected by replacing this with a continuous radius two-component Gaussian mixture. A Gauss-Hermite normalization error, in which mismatched quadrature nodes and weights deflated the fitted host-level mixing width σ_u by $\sqrt{2}$, was corrected by fixing the quadrature normalization. A radius error outlier, in which one FGK planet with a 68.9 $R_\oplus$ radius error distorted the calibration, was corrected by clipping relative radius errors to [1%, 30%]. A further set of issues comprised component degeneracy, in which free component widths allowed the optimizer to drive β_{age} toward zero, corrected by constraining component bounds to data measured values; objective scaling, which caused premature L-BFGS-B convergence at $\beta_{\mathrm{age}} = 0$, corrected by normalizing the objective by sample count; an anchor matching bug, in which substring matching produced spurious anchor correlations, corrected by exact post normalization string matching; and a premature "precise null" label, in which the initial FGK interpretation overstated the evidence against the literature effect, which was retracted and reported as inconclusive.

The third review pass identified one further defect in the injection machinery: the mapping from hostname to bootstrap index assigns a host drawn more than once a single shared random effect, taken from its last occurrence, rather than independent draws per copy. In a bootstrap of $n = 301$ hosts the expected number of distinct hosts drawn is ≈ 190, so ≈ 111 of 301 host-effect draws ($\approx 37\%$) go unused and effective host-effect diversity is reduced in all injected ensembles. The results reported here preserve the legacy behavior so that they remain comparable with the archived production run; the pipeline provides a corrected per-occurrence form (`independent_hosts=True` in `synth/synth_null`, `--independent-hosts` in `15_stratified_gate.py`) with an unchanged RNG stream. That corrected form was executed as a sensitivity check (Appendix A: M 200/arm, FGK 100/arm; production-scale confirmation M 1000/arm, FGK 400/arm): arm separations move by ≈ 0.1 null SD, from -0.19 to -0.36 (M) and from -1.01 to -0.83 (FGK) in the short runs, with the production-scale rerun giving -0.256 (M, 1000/arm) and -0.835 (FGK, 400/arm). Likelihood ratios and percentile placements are materially unchanged, and every verdict remains FAIL. The production legacy values are therefore retained as primary with the corrected reruns as the bounded sensitivity check (Table 4).

The reporting rule adopted after the audits is that analytical standard errors and p-values from the hierarchical mixture model are not treated as sufficient evidence on their own; the primary verdict is based on the empirical injection calibration.

4 Results

4.1 Selection Effects

Because the sample is drawn from a confirmed planet catalog rather than a uniform detection sample, we test directly whether the age axis is entangled with detectability.

Within the FGK Kepler only control, it is not. The Spearman correlation between $v_{\tan}$ and the modeled detection probability is -0.027 ($p = 0.21$), and between $v_{\tan}$ and the completeness weight is $+0.026$ ($p = 0.23$). Splitting at the median $v_{\tan}$, the median detection probability is 0.988 in the low half and 0.991 in the high half ($p = 0.80$). Whatever else is true of this sample, the age proxy is not a proxy for detectability.

The age axis is, however, entangled with measurement precision. Across the same median split, hosts in the high $v_{\tan}$ half are more distant (median 856 pc against 686 pc in the low half, $p = 3.8 \times 10^{-19}$), fainter (K_p 14.66 against 14.42, $p = 8.9 \times 10^{-8}$), and carry planets with larger relative radius errors (0.177 against 0.149, $p = 6.0 \times 10^{-6}$). Because the mixture model classifies planets softly through their radius uncertainties, larger measurement error pushes probability mass toward the overlap region between the two components, and it does so in a manner correlated with the age axis by construction. We test the consequences of this in Section 4.3.

The same tests applied to the cool sample find no entanglement of either kind. There, $\rho(v_{\tan}, P_{\det}) = -0.005$ ($p = 0.92$) and $\rho(v_{\tan}, w) = +0.006$ ($p = 0.91$), and the measurement precision confound does not reproduce: relative radius error is if anything weakly anticorrelated with the age proxy ($\rho = -0.090$, $p = 0.07$), and kinematically faster cool hosts are nearer rather than more distant (median 68 against 96 pc), the opposite of the FGK control. This is expected, because the cool sample is drawn from a shallower and more nearly volume limited population. The precision confound is therefore specific to the FGK control and is not an additional limitation of the cool-host analysis.

4.2 FGK Control: Calibrated Gate Result

Applied to the completeness-weighted FGK Kepler only control sample ($N = 2{,}135$ planets across 1,588 hosts; ESS $= 1{,}645$ after trimming at the 95th percentile), with 1,000 replicates per arm, the calibrated gate returns the values in the first column of Table 3.

Table 3. Calibrated gate results for all four configurations.

Quantity	FGK, pooled	FGK, stratified	Cool, pooled	Cool, DR25 subset
N (planets)	2,135	2,135	417	126
Hosts	1,588	1,588	301	82
ESS	1,645	1,645	140	117
Replicates per arm	1,000	400	1,000	500

Quantity	FGK, pooled	FGK, stratified	Cool, pooled	Cool, DR25 subset
Real β_{age} per IQR	-0.0296 ± 0.1061	-0.1482 ± 0.1019	$+0.0458$ ± 0.3210	$+0.3799$ ± 0.2962
Null injected median	-0.0186	$+0.0027$	-0.0513	-0.1345
Null injected SD	0.1169	0.1490	0.6720	0.3953
Signal injected median	-0.1365	-0.1142	-0.1812	-0.2438
Real percentile, null	45.6%	15.3%	55.3%	93.6%
Real percentile, signal	81.2%	40.8%	64.8%	96.8%
Likelihood ratio (signal:null)	0.68	1.74	0.93	0.56
Arm separation (null SD)	-1.01	-0.78	-0.19	-0.28
Verdict	FAIL	FAIL	FAIL	FAIL

Percentile placements are empirical Monte Carlo rankings of the real coefficient within the injected ensembles, not classical hypothesis test p-values. The likelihood ratio is the ratio of the two injected densities evaluated at the real coefficient, estimated by kernel density. Arm separation is the signed difference between signal and null injected medians in units of the null standard deviation. The FGK stratified column uses mixture components pinned to full sample values in both halves, so its pooled counterpart run under the same specification places the real coefficient at the 47.3rd percentile of null rather than the 45.6th, with an arm separation of -0.81; the two are consistent within Monte Carlo variation.

Table 4. Arm separations under sensitivity variants, in the signed convention (signal minus null) in units of the null standard deviation.

Variant	Cool / M	FGK	Replicates, M / FGK	Verdict
Production, primary	-0.19	-1.01	1,000 / 1,000	FAIL
Legacy host mapping, short run	-0.24	-0.94	200 / 100	FAIL
Corrected per occurrence, short run	-0.36	-0.83	200 / 100	FAIL
Corrected per occurrence, production scale	-0.256	-0.835	1,000 / 400	FAIL
RUWE and parallax S/N cut	-0.25	-0.84	200 / 200	FAIL

All values are signed (signal minus null)/null SD. The corrected per-occurrence form uses `independent_hosts=True` in the synthetic data generators; both a short run and a

production-scale confirmation are shown. Every verdict remains FAIL (inconclusive), so none of the sensitivity variants changes the conclusion.

The real coefficient lies at the 45.6th percentile of the null injected distribution while also lying at the 81.2nd percentile of the literature-sized signal distribution (Figure 4). The likelihood ratio between the hypotheses evaluated at the real coefficient is 0.68: the data mildly favor the null, by a margin conventionally regarded as not worth mentioning. The real result is therefore comfortably compatible with the null hypothesis and simultaneously compatible with a literature-sized effect. The 81.2nd percentile is a statement about the location of the real coefficient within the signal injected distribution, not evidence of detection. This result is a statement about the power of this pipeline under this control configuration; it is neither evidence for nor against FGK age evolution.

Signal recovery diagnostics show that the $\beta_{\mathrm{inj}} = -0.14$ injections are recovered with approximately 21% attenuation toward zero: the signal injected median of -0.1365 per IQR compares with an unattenuated expectation of $-0.14 \times 1.236 = -0.173$ per IQR. This is a known behavior of the mixture specification in this sample size regime and further motivates using the empirical injected distributions rather than analytical standard errors as the primary calibration. We note that the larger figure of approximately 35% quoted in Section 3.5 refers to the superseded fixed radius Bernoulli formulation, not to the mixture specification adopted here.

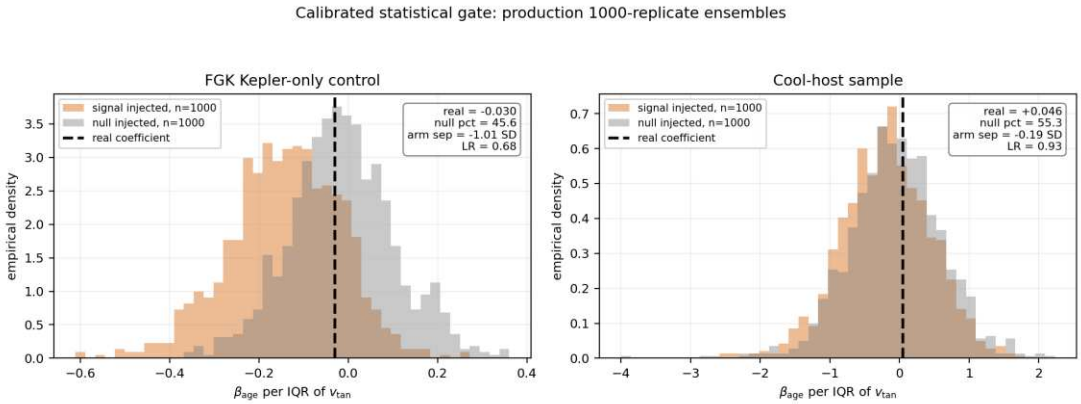

Figure 4. Calibrated statistical gate applied to the FGK control (left) and the M dwarf sample (right). Gray is the null injected ensemble, orange the literature-sized signal injected ensemble, and the dashed vertical line the real coefficient. The signed FGK arms are separated by -1.01 null standard deviations and still overlap enough to produce an inconclusive verdict; the signed M dwarf arms are separated by -0.19 and are for practical purposes the same distribution. Note the difference in horizontal scale between the panels.

4.3 A Stratified Estimator That Fails the Gate

The measurement precision confound identified in Section 4.1 suggests that the pooled fit may be diluted by a nuisance term correlated with the age axis. We tested this by refitting

the production model within halves of the sample, with the mixture components pinned to full sample values so that splitting cannot move them. The results are given in Table 5.

Table 5. Slice refits of β_{age} in the FGK control.

Slice	N	β_{age} per IQR
Pooled baseline	2,135	-0.0296 ± 0.1061
K_p bright half	1,069	-0.1356 ± 0.1287
K_p faint half	1,066	-0.1695 ± 0.1668
Distance, near half	1,049	-0.0971 ± 0.1214
Distance, far half	1,044	-0.2186 ± 0.1884
Precision, precise half	917	-0.0761 ± 0.1198
Precision, blurry half	916	$+0.1107 \pm 0.2638$

Mixture components are pinned to full sample values in every row, so the splits cannot move them. These partitions were selected post hoc and are adjudicated by the calibrated gate in Table 3.

Both halves of every partition lie further from zero than the pooled fit, and the blurry precision half is the only slice with the opposite sign. Inverse variance pooling the brightness partition gives -0.148 ± 0.102, which is 85% of the literature-sized value and a nominal z of -1.46.

Because this stratification was selected post hoc, we ran the stratified estimator through the identical gate, with 400 replicates per arm (Table 3, second column). It fails. The real stratified coefficient sits at the 15.3rd percentile of its own null injected ensemble, with a binomial 95% interval of [11.7%, 18.8%] and a likelihood ratio of 1.74.

The mechanism is instructive. The real data's two halves agree to within 0.034, but the median gap between halves in the null ensemble is 0.166, and only 10.2% of null replicates agree that closely (12.0% under the signal arm). Such concordance is a one in ten fluctuation under *both* hypotheses and therefore carries no information about which is true. It is also what produced the unusually tight analytical interval: the real stratified standard error of 0.102 sits at the 2.8th percentile of the null arm's own standard errors, which converts a nominal z of -1.46 into an empirical z of -1.01 against the null ensemble.

We report this as a negative control. A plausible, mechanistically motivated estimator produced an apparently encouraging result that the calibration refused to certify. This demonstrates specificity: the gate can reject a spurious estimator. It does not demonstrate sensitivity, because no real fit in this study passes. The injection arms themselves show sensitivity in the expected sense: their signed separation grows from -0.059 at $\beta_{\mathrm{inj}} = -0.07$ to -0.379 at -0.2206 (complete 150-replicate curve, archived as described in the Data and Reproducibility Availability section), so the procedure responds to a larger injected effect even though the real coefficient remains null-like. We do not

claim that this decision protocol is a stronger evidentiary standard than the completeness-corrected injection validation used by Gaidos et al. (2024); it answers a different question.

4.4 Cool Hosts: Calibrated Gate Result

The completeness-weighted radius distributions behind both samples are shown in Figure 3. Applied to the cool sample ($N = 417$ planets across 301 hosts; ESS $= 140$) with 1,000 replicates per arm, the real coefficient is $+0.0458 \pm 0.3210$ per IQR, sitting at the 55.3rd percentile of the null injected distribution with a likelihood ratio of 0.93 (Table 3, third column). The sign is positive, that is, opposite to the direction predicted by atmospheric mass loss, but at $z = +0.14$ against the empirical null this should not be interpreted.

The verdict is driven by the placement of the real coefficient in the null ensemble, which is independent of the injected effect size. Arm separation is a diagnostic of available power rather than the decision criterion. The signal and null injected medians differ by -0.19 null standard deviations, against -1.01 for the FGK control (Figure 4). The two hypotheses generate distributions that are, for practical purposes, the same distribution. Under the tested mixture model and calibrated gate, this sample cannot distinguish a literature-sized age effect from no effect. Within this tested specification the failure reflects limited dataset information; we make no claim about all possible estimators.

Stratification was attempted for the M dwarf sample and abandoned: 55% of stratified fits failed to converge under the brightness partition (57% under the signal arm) and 65% under the distance partition. With ESS 140 splitting into roughly 70 effective planets per half, the mixture model does not identify. The surviving fits form a convergence selected subsample whose percentile placements would be biased by an unknown amount, and we therefore do not report them.

As a sensitivity check we repeated the gate on the DR25 CDPP subset (Table 3, fourth column). The point estimate of $+0.3799 \pm 0.2962$ is consistent within its large uncertainty with the full sample fit of $+0.0458 \pm 0.3210$, despite discarding 70% of the planets. Because the subset hosts are hotter and kinematically cooler than the full sample, this comparison suggests that the proxy weighted planets contribute variance, but a variance-only reading is confounded by the different host population. The null distribution is in fact tighter (SD 0.395 against 0.672) and the arm separation slightly better (-0.28 against -0.19), so the cleaner subset is marginally closer to being able to answer the question than the full sample despite having a third as many planets. The verdict is unchanged, and the subset is not representative of the target population (Section 3.1).

4.5 Period Mixing

During the FGK control analysis, a significant apparent wrong-sign signal appeared in binned tests: a completeness-weighted two proportion test at a fixed 1.88 $R_\oplus$ classification showed the sub-Neptune fraction higher among kinematically old (high $v_{\tan}$) hosts by $+5.01$ percentage points (se 2.31 pp, $z = +2.16$, $p = 0.0304$), opposite to the expected direction from atmospheric mass loss. The hierarchical model, however,

reported $\beta \approx 0$. Rather than choosing the more convenient result, we investigated the discrepancy systematically. The binned two-proportion excess and the logistic average marginal effect are different estimands; an earlier statement of "$+6.29$ pp with $z = 2.16$" conflated the logistic point estimate with the binned z-statistic.

We first tested whether metallicity confounding could explain the artifact, since kinematically hot stars tend to be metal poor and metallicity affects planet compositions. Stratifying the sample into [Fe/H] terciles gave z-statistics of $+0.49$, $+1.55$, and $+0.62$ for the metal poor, solar, and metal rich terciles respectively, all non-significant. The inverse variance pooled estimate was $\Delta_{\mathrm{SN}} = +0.042 \pm 0.028$ ($z = +1.50$, $p = 0.134$), with heterogeneity $Q = 0.69$ (χ^2 $p = 0.71$). A full power weighted logistic regression with [Fe/H] as the only covariate (complete case $N = 1{,}479$) yielded an average marginal effect of $+6.27 \pm 2.90$ percentage points ($p = 0.030$) and $\beta_{[\mathrm{Fe/H}]} = -0.89 \pm 0.50$ per dex ($p = 0.07$). The metallicity hypothesis was therefore not supported as the primary explanation. We note that the complete case restriction is unlikely to be missing at random, since [Fe/H] availability tracks host brightness and follow up history.

Adding $\log_{10}(P)$ as a covariate reduced the average marginal effect to $+4.14 \pm 3.54$ percentage points ($z = +1.17$, $p = 0.241$) in the delta-method computation archived with the analysis outputs; the model-based regression error of ± 2.61 pp ($z = 1.594$, $p = 0.11$) is retained in the first-party ablation record, and the qualitative conclusion is unchanged under either convention. The coefficient on $\log_{10}P$ is $+2.38 \pm 0.17$. Kinematically old hosts have slightly longer period planets in this sample (median 10.69 d versus 10.05 d), while the sub-Neptune fraction rises steeply with period against a fixed 1.88 $R_\oplus$ classification (Figure 5). The resulting composition difference can therefore generate a spurious wrong-sign excess among high $v_{\tan}$ hosts when period is ignored.

We describe this as substantial attenuation rather than identification of a complete artifact: conditioning on period reduces the point estimate by roughly a third and removes nominal significance, but does not eliminate it, and the two estimates lie within each other's uncertainties. Period mixing is a contributing mechanism, not a demonstrated full explanation. Within four equal-width $\log_{10}P$ bins the old-minus-young sub-Neptune excesses are $+0.96$, $+1.15$, $+6.50$ ($p = 0.060$), and -0.80 percentage points; none of the four bins is significant at the 5% level. A host-level permutation control with 2,000 permutations places the observed $+4.14$ pp at one-sided $p = 0.045$ and two-sided $p = 0.104$, so the residual is weak and borderline rather than a robust detection.

The period trend itself is robust to the weighting. Panel (b) of Figure 5 shows the sub-Neptune fraction both unweighted and completeness-weighted; the trend is present in both, and weighting lowers the fraction in every bin rather than inflating the long period end. The mechanism is therefore not an artifact of inverse detection weighting at long periods, which was a live possibility given the size of the weights in that regime.

The hierarchical model does not display this artifact because it includes log period as a covariate from the outset. The broader implication is methodological rather than universal:

archival demographic comparisons that group stars by age, metallicity, kinematics, or another variable correlated with period should explicitly control for orbital period or verify comparable period distributions before interpreting differences in sub-Neptune fraction.

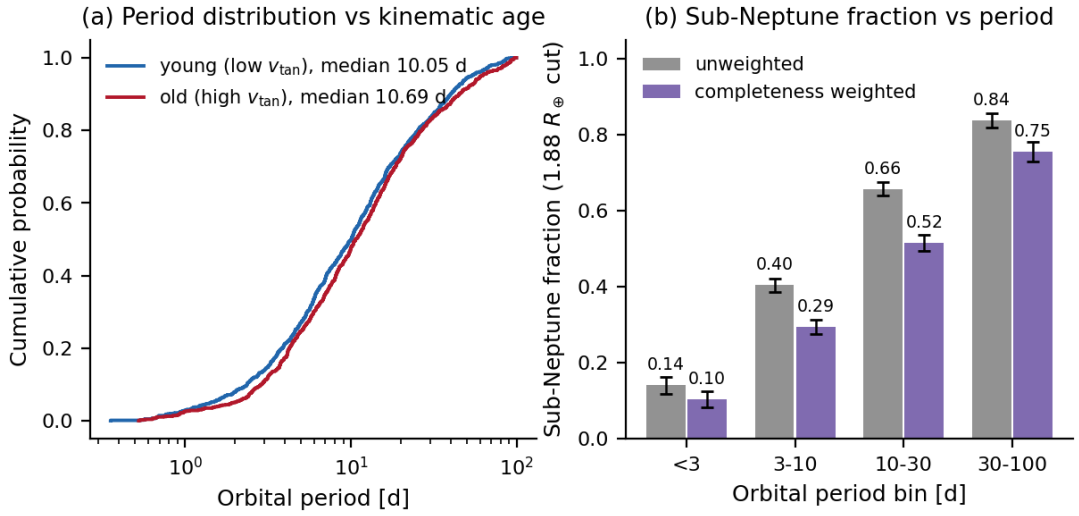

Figure 5. Period mixing in the FGK control. (a) Kinematically older hosts carry slightly longer period planets (medians 10.69 d against 10.05 d). (b) The sub-Neptune fraction rises steeply with orbital period under a fixed 1.88 $R_\oplus$ classification. The trend is present both unweighted and completeness-weighted, so it is not an artifact of inverse detection weighting; weighting lowers the fraction in every bin. Error bars use the effective sample size in each bin.

4.6 Power Analysis

We quantify the statistical path to resolving the M dwarf question with a 15 cell Monte Carlo power grid. The grid bootstraps hosts and covariates from the real M dwarf sample at five scale factors ($0.5\times$, $1\times$, $2\times$, $4\times$, $8\times$ current size) and three injected effect sizes (null: $\beta_{\rm inj} = 0$; literature-sized: $\beta_{\rm inj} = -0.14$ per SD; strong: $\beta_{\rm inj} = -0.28$ per SD). Planet radii are drawn from class conditional distributions matched to the observed cool-host modes; completeness weights are applied; each realization is fitted with the mixture likelihood model. The injected component values are close to but not identical to the component parameters pinned in the refit, and the identical specification is used for both injected arms, so this is a property of the calibration rather than a source of differential bias between arms. There are 500 replicates in every cell of the production grid; the raw replicate level results are the archived ground truth, and the tabulated summaries reproduce them exactly.

Table 6. Power grid results for the M dwarf sample.

Scale	$N_{\rm pl}$	$N_{\rm eff}$	Null FPR	Lit. det.	Strong det.
$0.5\times$	~208	~70	7.2%	9.0%	16.2%

Scale	N_{pl}	N_{eff}	Null FPR	Lit. det.	Strong det.
$1\times$ (current)	~417	~140	8.2%	9.4%	27.0%
$2\times$	~834	~279	5.2%	13.6%	38.8%
$4\times$	~1,667	~558	5.6%	18.8%	50.2%
$8\times$	~3,337	~1,118	7.2%	23.2%	60.6%

N_{pl} is a planet count; N_{eff} is the corresponding effective sample size at the 33% ESS retention measured for this sample (Table 2). Null FPR is the false positive rate under $\beta_{\mathrm{inj}} = 0$; Lit. det. and Strong det. are detection rates under the literature-sized and strong injected effects respectively. All cells use 500 replicates.

The earlier 120 replicate grid showed an apparently non-monotonic null false positive rate (11.7% at $1\times$, 2.5% at $4\times$, 7.5% at $8\times$). The 500 replicate grid resolves this: the false positive rate is 7.2%, 8.2%, 5.2%, 5.6%, and 7.2% across the five scales, so the previous $4\times$–$8\times$ contrast was finite Monte Carlo fluctuation and there is no special $8\times$ anomaly (Fisher's exact test of the $8\times$ cell against any other cell gives $p \geq 0.24$). Pooled across all five null cells the false positive rate is $167/2500 = 6.68\%$, significantly above nominal (binomial $p \approx 10^{-4}$), with a mild, approximately scale independent anti-conservative tendency that also affects the largest cells. For historical comparison only, the superseded 120-replicate grid gave a null false positive range of 2.5%–11.7% and a coverage range of 83%–98%; those figures are not the values adopted here.

This inflation, and the corresponding under coverage of the analytical intervals (pooled empirical coverage $6850/7500 = 91.3\%$ against the nominal 95%, with a per cell range of 85.2%–95.4% and the worst cell, $8\times$ strong, at 85.2%), is a property of the analytical confidence interval and detection rule used inside the power grid diagnostics. It is consistent with the documented tendency of the analytical sandwich standard errors to be somewhat too small for this mixture model and class overlap structure, and it does not disappear at the largest simulated scale. Per-cell single-sample high-replicate ranges are 5.2%–8.2% for the false positive rate and 85.2%–95.4% for coverage. It does not affect the primary calibrated gate, whose verdict is based on empirical injected distributions rather than analytical intervals; we therefore distinguish the empirical calibration of the gate from the analytical interval performance inside the power diagnostics. Recomputing detection against the empirical null 2.5% quantile at each scale (one-sided, correct sign; see the Data and Reproducibility Availability section, Wilson 95% intervals) gives lower values than the tabulated analytical rates, which strengthens rather than weakens the conclusion: at $1\times$, 7.2% [5.2%, 9.8%] instead of 9.4% for the literature-sized effect (20.0% instead of 27.0% for the strong effect); at $8\times$, 19.8% [16.5%, 23.5%] instead of 23.2% (55.8% instead of 60.6%). The full 15-cell size-adjusted table is archived with the analysis outputs (the Data and Reproducibility Availability section).

At the current scale (~417 planets), the literature-sized effect is detected only 9.4% of the time. Detection rises monotonically with scale, reaching 23.2% (95% empirical range 19.6%–27.2%) at $8\times$ (~3,337 planets), while a strong ($2\times$ literature) effect reaches 60.6%

(95% empirical range 56.2%–64.9%). These are empirical simulation results under the current modeling assumptions, not universal future survey predictions. Because the upsampling procedure resamples the current empirical covariate distribution, it cannot capture different noise properties, stellar population mixes, or survey selection functions in future datasets.

We stress that the scale factors are planet counts. At the 33% ESS retention measured for this sample, an 8 × planet inventory corresponds to roughly 1,100 effective planets rather than 3,337 (Table 6), and the horizontal axis of Figure 6 is annotated accordingly. Improvements in per-star completeness products for TESS observed M dwarfs would therefore raise effective sample size at no observational cost.

The key conclusion is conservative: even the largest tested scale remains far below conventional 80% power for a literature-sized effect (Figure 6). The 8 × scale should be interpreted as a floor or intermediate benchmark, not as a sufficient target.

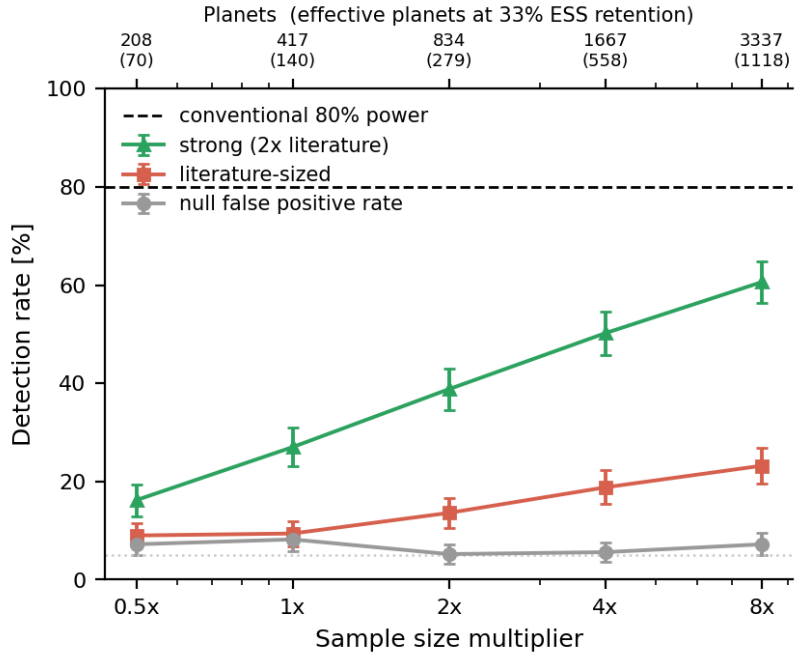

Figure 6. Power versus sample size scale for the M dwarf analysis, from the 500 replicate grid. The dashed horizontal line marks the conventional 80% power reference. No tested scale reaches that threshold for the literature-sized effect. The upper axis gives planet counts and, in parentheses, the corresponding effective planet counts at the sample's 33% ESS retention. Error bars are binomial 95% intervals.

Extrapolation.

For orientation only, a simple Gaussian extrapolation using $\sigma \propto N^{-1/2}$ applied to the simulated 8 × power ($p_8 = 0.232$, giving a standardized effect $\Phi^{-1}(p_8) + 1.96 = 1.23$ against the conventional 80% threshold of 2.80) places conventional 80% power at $N_{80} \approx$

1.7×10^4 planets. However, the grid itself does not exhibit $N^{-1/2}$ scaling: because the bootstrap repeatedly resamples the same ~301 M dwarf hosts, the median replicate standard errors scale with the $8\times$ cell as 2.37, 1.73, 1.36, 1.12, and 1.00 from $0.5\times$ to $8\times$, whereas fully independent hosts would predict 4.00, 2.83, 2.00, 1.41, and 1.00. The measured exponent is $\sigma \propto N^{-0.31}$, and applying it instead of $N^{-1/2}$ gives $N_{80} \approx 4.7 \times 10^4$. We report both as a bracket rather than quoting the more optimistic value alone.

Host covariate diversity therefore saturates at large scale, which may make the simulated scaling pessimistic relative to a future sample with genuinely new hosts; unmodeled differences in future survey selection and noise could move the requirement in the opposite direction. The extrapolated scale should accordingly be read as an order of magnitude illustration rather than a forecast with a sharp numerical threshold.

5 Discussion

5.1 Implications for the M Dwarf Radius Valley

The central result is the inconclusive verdict itself: under the tested mixture model and calibrated gate, the current cool-star inventory cannot discriminate a literature-sized age effect from a null effect. The signed arm separation of -0.19 null standard deviations quantifies the power available for that adjudication. The inconclusive FGK control provides the calibration for this statement: a sample with twelve times the effective information reaches a signed arm separation of only -1.01 and still returns FAIL.

This has direct implications for the formation based hypothesis of Gillis et al. (2026). Their proposal that the radius valley around mid to late M dwarfs is absent because it was never formed rather than because it evolved away makes a testable prediction about age dependence. Our analysis shows that the prediction remains untested, not confirmed or refuted. The absence of a detected age signal is consistent both with a formation only scenario and with an attenuated atmospheric evolution signal below the current detection threshold.

This conclusion must be read alongside Gaidos et al. (2024), who did find an age-dependent sub-Neptune to super-Earth ratio in this same temperature range. Their ratio falls from 2.38 to 0.58 across a median split near 3.86 Gyr with KS $p = 0.08$, and their Section 2.2 and 2.3 apply Kepler DR25 detection-efficiency corrections and synthetic injection and recovery validation. Their Appendix A further shows that planets are at most about 10 percent less detectable around younger active stars, far too small to explain the factor-of-four change in the number ratio (the sub-Neptune fraction itself halves, from 0.704 to 0.367). That result is a marginal completeness-corrected contrast in the cool-host regime: conventionally suggestive at $p = 0.08$, but not meeting the 5 percent criterion our gate applies to its own real coefficient. Our verdict does not overturn it. The 65 finite-age TIME Table matches supply a direct empirical check, reported in Section 2.4: $v_{\tan}$ correlates with published gyrochronological age at $\rho = +0.372$ ($p = 0.0023$), so the ordinal

kinematic axis does carry age information in this temperature range even though its zero point is unvalidated. Because those ages fall outside the Angus et al. (2019) validity domain and we have not run the gate on them, sample size and age precision remain entangled as explanations for the inconclusive verdict. The distinction is the instrument: their test asks whether an observed ratio changes with age after completeness correction, while ours asks whether the present archive can discriminate a literature-sized effect from zero under an injected null and signal decision rule. The two answers are compatible if the underlying effect is larger than the value we inject, if the kinematic proxy dilutes the age baseline, or if the different radius classifications target different populations. A further alternative is that the population near 1.25 $R_{\oplus}$ reflects a water-condensation boundary rather than an atmospheric-loss valley (Luque and Pallé 2022), in which case an age trend in a radius-based classification need not trace envelope loss. We therefore do not claim that M dwarf age demographics are untested; we claim that a calibrated detectability decision on this heterogeneous catalog is inconclusive, and we identify the rotation-aged, completeness-controlled subsample as the decisive next test.

We remain mechanism agnostic. Our result does not discriminate among photoevaporation (Owen and Wu 2017), core-powered mass loss (Ginzburg et al. 2018), pebble accretion and water-world formation (Lopez and Rice 2018), steam-world scenarios, or a primordial valley. Each remains viable for this sample, and the gate is not constructed to favor any one of them.

Three limitations compound and should be read together. First, population definition: the present M dwarf sample uses $T_{\rm eff} < 4200$ K and therefore spans a broader, heterogeneous cool-star regime than the approximately M3 and later population associated with the valley disappearance result, and cannot test that M3+ population directly. Subdivision by spectral type is an important next step, but the present sample is far too small for that exercise to carry inferential weight. Second, age information: only 12 of 303 M dwarf hosts yield usable gyrochronological ages within our pipeline's Angus+19 domain, where $\rho = 0.25$ at $p = 0.43$ ($n = 12$); an external TIME Table crossmatch validates ordinal ordering ($\rho = 0.37, p = 0.0023, n = 65$) but supplies no absolute zero point (Section 2.4). Third, effective sample size: 33% ESS retention means the nominal 417 planets behave as 140, and 70% of the weights rest on a proxy detection model rather than measured photometric noise.

5.2 Effect Size Versus Power

Increasing the injected effect does not rescue the gate for this sample. We ran a complete power versus effect curve with ten injection points and 150 replicates per arm, using a single null arm of 150 draws (Figure 7). The points are $\beta_{\rm inj} = 0, -0.07, -0.105$ (Sandoval et al. 2021), -0.1419 (Kamulali et al. 2026), $-0.2206, -0.3089$ (Berger et al. 2020), $-0.4412, -0.60, -0.8824$ on Gaidos's true-age scale, and -1.20, listed in monotonic order of increasing injected effect. The two ρ-scaled values of the Gaidos effect are illustrative bounds on imperfect age information and lack a pipeline-domain calibration for this sample, since the cool-host $\rho = 0.25$ is not significant, although the external TIME Table

crossmatch does establish ordinal ordering at $\rho = 0.37$ ($p = 0.0023$, $n = 65$). We do not adopt the FGK attenuation factor as a primary correction; a factor of 0.5 is shown here only as an illustrative bound. Because the real cool-host coefficient ($+0.0458$ per IQR) sits at the 54.7th percentile of this curve's 150-replicate null arm (55.3rd in the production 1,000-replicate gate of Section 4.4), it cannot satisfy the one-sided distinguishability condition for any injection, and the verdict is FAIL at every point. Arm separation grows monotonically from 0.00 at the null to -0.225 at the primary effect, -0.846 at the $\rho = 0.5$ bound, and -1.575 at the Gaidos-sized effect. The detection rate rises from 2.7 percent under the null to 42.0 percent at the Gaidos-sized injection, and reaches 57.3 percent only at the largest, deliberately extreme point. The gate therefore responds strongly to a larger signal: a Gaidos-sized effect produces clearly separated injected ensembles (arm separation -1.58 null SD), although the two-condition gate detects it in only 42.0 percent of replicates; the real coefficient remains null-like. The inconclusive verdict is driven by the data rather than by an instrument that cannot respond. The limitation is the information content of the cool-host sample and its age axis, not a particular choice of effect size.

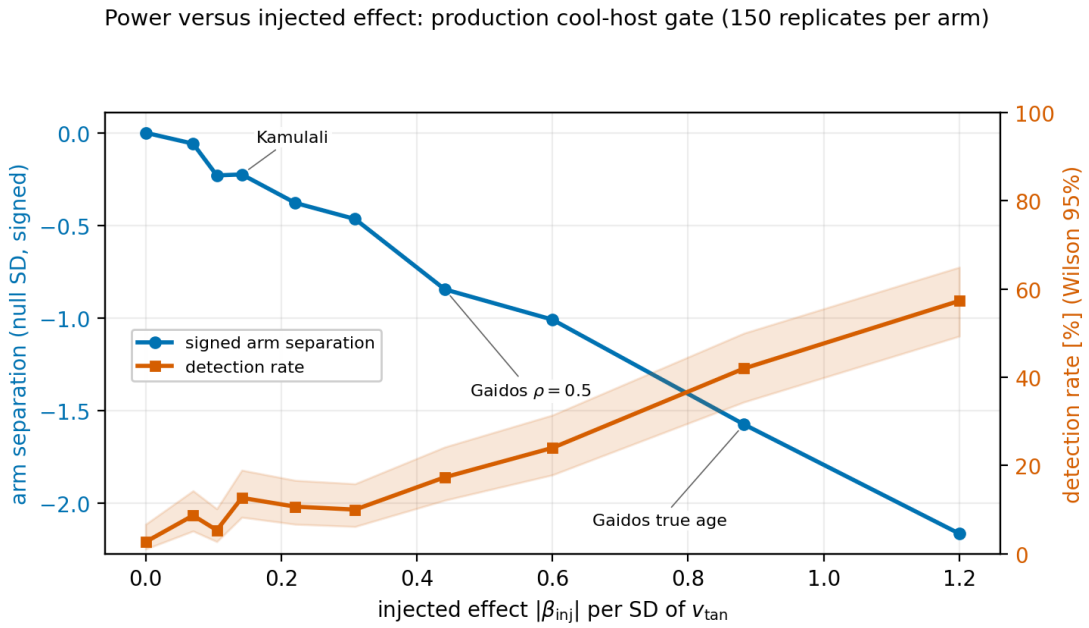

Figure 7. Power versus injected effect for the calibrated cool-host gate, from ten points with 150 replicates per arm. The signed arm separation (signal minus null medians in null standard deviations) grows monotonically with the injected coefficient, and the detection rate rises from 2.7 percent under the null to 42.0 percent at the Gaidos-sized effect of -0.8824 per SD. Every point returns FAIL because the real coefficient's null placement is independent of the injection. Labeled points are Kamulali (0.1419), the $\rho = 0.5$ Gaidos bound (0.4412), and the Gaidos true-age value (0.8824); the unlabeled points are the null, half-effect, Sandoval, Berger, $\rho = 0.25$ bound, and the two extreme injections. Error bars are Wilson 95 percent intervals.

5.3 Period Mixing as a General Methodological Concern

The period mixing effect identified in Section 4.5 is not specific to this choice of age proxy. In this dataset, different stellar subgroups have slightly different orbital period distributions, and the sub-Neptune fraction changes strongly with period. A naive binned comparison can therefore attribute a period driven demographic difference to the grouping variable.

The appropriate lesson is narrow but broadly applicable: archival radius valley demographic comparisons should explicitly condition on orbital period, or verify that comparison groups have sufficiently similar period distributions, before interpreting differences in sub-Neptune fraction. The present analysis identifies this vulnerability in the dataset; it does not claim that the same magnitude of bias must occur in every archival study.

A period-dependent classification is the direct remedy. Greco et al. (2026) fit the valley locus as $\log_{10}[R_{\rm gap}(P)/R_\oplus] = m\log_{10}(P/{\rm d}) + q$ with $m \simeq -0.096$ and $q \simeq 0.329$ on FGK hosts, giving $R_{\rm gap}(10{\rm d}) \simeq 1.71R_\oplus$, and classify each planet relative to the boundary at its own period. Because cool-star period slopes may be flatter or of opposite sign (Cloutier and Menou 2020; Gaidos et al. 2024), we would not import that slope directly. The appropriate next step is to fit m on the cool-host sample or bracket it over $m \in [-0.10,0]$, rescale radii as $R_p' = R_p(P/10{\rm d})^{-m}$, and repeat the binned age test. Across this bracket the adjusted average marginal effect is 4.45–5.25 pp with $p > 0.12$ in every case, and we label these values upper bounds because the data-driven fit hits the search boundary and m is not identifiable in this sample. Reporting the unadjusted and period-adjusted estimates side by side remains the honest interim treatment; the adjusted value is an upper bound rather than a detection.

A companion blending budget reinforces this reading. Bergsten et al. (2026) find that unresolved companions raise small-planet occurrence by factors of 1.08 to 1.19, with a 6–24% differential that is larger for super-Earths and at longer periods. Applied as an age-differential detectability contrast, this is both too small and mis-targeted to account for the $+4.14$ pp period-adjusted residual: explaining it would require an implausible age-differential companion incidence: even the lower end is 43 percentage points, above the roughly 25 percentage point observed-to-field contrast in that work, and the formal upper estimate exceeds 100 percent and is therefore unphysical. We did not perform an end-to-end crossmatch of our hosts to the Gaia bound-pair catalog, so no binary fraction for our sample is claimed and this remains a literature-based order-of-magnitude budget.

5.4 The Path to Resolution

Planet count alone will not resolve this question under the current modeling assumptions. Three factors bind simultaneously.

Independent host count matters because stellar age is a host-level quantity, and our own bootstrap demonstrates that host covariate diversity saturates once the same hosts are resampled. Per-host age precision matters because the kinematic proxy carries roughly

half the signal even where it is validated, and lacks a pipeline-domain absolute calibration where it is most needed (Section 2.4). Effective sample size matters because completeness weighting currently discards two thirds of the M dwarf sample's nominal information, much of it because the weights rest on a proxy detection model.

An approximately eightfold expansion may be relevant to the yield territory of PLATO and future terrestrial planet finding observatories, but even that scale would leave the question statistically unresolved under the current modeling assumptions. PLATO is expected to deliver asteroseismic ages at roughly 10% precision for bright hosts (Rauer et al. 2025), and TESS observations of young clusters offer an independent age calibration (Rogers 2025). Both are relief on the age axis rather than a substitute for more independent hosts. Materially better stellar age measurements could improve power without a proportional increase in planet count, and improved per-star completeness products for TESS observed M dwarfs would improve it further at no observational cost at all.

An attempted validation using isochrone based ages in place of the kinematic proxy encountered unresolved data and calibration issues and was not completed. This remains a useful direction for future investigation, especially if homogeneous stellar parameters and age estimates become available for a larger M dwarf sample.

5.5 Audit Trail as Methodological Practice

The audit history is a substantive part of the study because several identified issues materially affected the inference. The mixture identifiability correction, quadrature normalization, radius error correction, component constraints, period mixing diagnosis, host clustering checks, and the injection machinery defect described in Section 3.5 all reduced specific routes by which the analysis could have produced a misleading result. Presenting these corrections openly is intended to distinguish a calibrated inconclusive result from a nominally precise result whose assumptions were not stress tested. The stratified estimator of Section 4.3 belongs to the same category: it is reported precisely because it failed.

5.6 Robustness to Within-Host Clustering

Planets sharing a host star are not independent draws, so ignoring host-level dependence could in principle understate the uncertainty on β_{age}. The production specification addresses this with host-level cluster robust sandwich standard errors and host preserving injection ensembles. We additionally re-ran the primary FGK gate using one planet per host, selecting the largest-radius planet per host ($N = 2{,}135 \rightarrow 1{,}588$ planets); a uniformly random selection per host gives the same result to within Monte Carlo variation.

The coefficient changes from -0.030 ± 0.106 to -0.041 ± 0.123 per IQR of $v_{\tan}$, an SE inflation of approximately 16%, while the percentile placement remains essentially unchanged within the corresponding calibration ensemble: 43.3% in the 120 replicate robustness check's own null ensemble versus 45.6% in the 1,000 replicate primary gate. We note that these two figures come from ensembles of different size and are therefore

indicative rather than a like for like comparison. The verdict remains FAIL (inconclusive). These checks indicate that within host clustering does not materially alter the primary FGK inference under the tested specifications. They do not prove that no richer hierarchical structure could matter in future analyses.

6 Conclusion

We set out to test whether age evolution of the radius valley extends to M dwarfs, motivated by the recent finding that the valley disappears around approximately M3 and later dwarfs. The study is scoped along three axes: cool hosts with $T_{\rm eff} < 4200$ K, a quantitative physics versus power adjudication, and an explicit PASS/FAIL injection gate. We developed that gate by injecting null and literature-sized synthetic effects into the same noise structure and requiring the real data to discriminate between competing hypotheses.

In the FGK Kepler only control, the real age proxy coefficient (-0.0296 ± 0.1061 per IQR) lies at the 45.6th percentile of the null injected distribution and the 81.2nd percentile of the literature-sized signal distribution, with a likelihood ratio of 0.68 between the hypotheses and a signed arm separation of -1.01 null standard deviations. The control therefore cannot independently confirm the literature effect, but neither does it provide evidence against it.

For cool hosts the inconclusive verdict is more clearly power-limited: the two hypotheses generate nearly coincident distributions, separated by -0.19 null standard deviations against -1.01 for the FGK control, with a likelihood ratio of 0.93 at the real value. The question is unanswered by this sample under the tested mixture model and gate; we do not claim in-principle unanswerability with other estimators or data.

We validate the kinematic age proxy against gyrochronological ages, obtaining $\rho = 0.50$ ($n = 79$, $p = 3.0 \times 10^{-6}$) for FGK hosts but only $\rho = 0.25$ ($n = 12$, $p = 0.43$) for M dwarfs. An independent crossmatch to the Gaidos et al. (2023) TIME Table raises the cool-host validation set to 65 finite-age hosts and yields $\rho = 0.372$ ($p = 0.0023$), showing that the kinematic ordering carries age information in this regime. We show that the age axis is not entangled with detectability but is entangled with measurement precision, and we report a stratified estimator that appears to recover a literature-sized effect yet fails the same calibrated gate, with the apparent improvement traceable to a between half concordance that occurs in 10% of null replicates. We identify period mixing as a contributing mechanism behind a wrong-sign binned signal and show that the underlying period trend is robust to the completeness weighting. We report that the M dwarf sample retains only 33% of its nominal information under completeness weighting, so it behaves as a 140 planet sample.

The power grid provides a negative but useful planning result. Even at $8\times$ the current M dwarf inventory (~3,337 planets, or ~1,100 effective planets), a literature-sized effect is detected in only 23.2% of simulated realizations, and no tested scale reaches

conventional 80% power. A Gaussian independent host extrapolation places an order of magnitude scale for 80% power near 1.7×10^4 planets under $N^{-1/2}$ scaling, or near 4.7×10^4 under the $N^{-0.31}$ scaling actually measured in the grid; neither is an empirically demonstrated requirement. The central conclusion is that the present archival inventory cannot adjudicate the M dwarf age evolution question under the tested mixture model and gate, and that resolving it will require substantially more independent M dwarf systems, materially better age information, and better per-star completeness products, rather than a larger planet count alone.

Data and Reproducibility Availability

The NASA Exoplanet Archive planetary systems table, Gaia DR3, the Kepler DR25 stellar table, and the McQuillan et al. (2014) rotation catalog are publicly available. The full TIME Table of Gaidos et al. (2023) is archived at Zenodo DOI 10.5281/zenodo.7578269. The complete analysis pipeline, including archive queries, the hierarchical mixture model, completeness weighting, the calibrated gate, the power grid, the age proxy validation outputs (`proxy_validation_fgk_recomputed.json` and `timetable_spearman.json`), the effect-size and period-classifier records (`effect_size_derivation.json` and `period_classifier_results.csv`), the full effect-versus-power curve (`power_vs_effect_summary_FINAL.json`), the stratified gate (`15_stratified_gate.py`), the arm-delta sensitivity runner (`appendix_A_arm_delta.py`), the beta-distribution runner (`appendix_B_beta_dist.py`), the RUWE sensitivity runner (`appendix_C_ruwe.py`) with Gaia TAP fetchers (`fetch_gaia_quality_v2.py`), the likelihood-ratio check (`compute_gate_lr.py`), the size-adjusted power script (`size_adjusted_power.py`), the production-scale confirmation (`confirm_independent_production.py`), and the high-replicate orchestration script (`colab_highrep_gate_fast_v3.py`), is archived at https://github.com/zabdax/radius-valley-age-evolution. The `REVISED_PACKAGE/code/` directory ships the pipeline files used for the revision appendices; the original diagnostic and figure-generation scripts are retained in the main repository.

Archived releases: v2.0.3 at https://doi.org/10.5281/zenodo.24444755, which archives the analysis pipeline, scripts, results, and figures for this revision; the earlier v2.0.1 release at https://doi.org/10.5281/zenodo.23007310 preserves the superseded preliminary material. A bare GitHub URL is not a persistent archive.

For the high-replicate Monte Carlo experiments reported here, the raw replicate level CSVs are the archived ground truth, and the JSON summary files agree with them exactly; both are provided as supplementary materials. The original FGK production run was executed on Google Colab (CPU) with serial synthetic data generation and a two-worker process pool for fitting; the run manifest records an elapsed time of 6.0 hours and no skipped grid cells. The stratified and M dwarf gates use a reimplemented synthetic data generator with an index-based host gather in place of the original per-host concatenation, which reduces the cost per replicate by more than an order of magnitude; the script provides a `--verify` mode that asserts bit-level equivalence of the random stream against the original

implementation, and per replicate seed spawning so that results are independent of the number of parallel workers and of any resume point.

The model specification (working independence mixture likelihood, host-level cluster robust sandwich standard errors, and host bootstrap injection ensembles with one shared host effect per host, $\sigma_u = 1.5$) is defined by the archived pipeline code.

Arm separations quoted in these appendices come from shorter sensitivity ensembles and are not directly comparable to the production values in Table 3; the production legacy arms are -0.19 (M) and -1.01 (FGK), and all verdicts are unchanged across ensembles. Appendix A reports the independent-host sensitivity rerun (pooled estimator; M 200 replicates/arm with arms -0.24 legacy vs -0.36 independent; FGK 100 replicates/arm with -0.94 vs -0.83; verdicts unchanged; MCSE on each arm separation ≈ 0.12–0.19 null SD, so a shift of about 0.1 is noise-consistent), confirmed at production scale with the corrected generator (M 1000/arm arm -0.256, FGK 400/arm arm -0.835). Appendix B reports the beta-distribution injection (Kamulali uncertainties propagated; injected median ≈ -0.14, SD ≈ 0.16; M correct-sign detection 13.3% legacy / 11.3% independent). Appendix C reports the RUWE/parallax-S/N sensitivity gate (2,424 hosts re-queried; M arm -0.25, FGK arm -0.84; verdicts unchanged). Raw replicate CSVs (`arm_delta_all_raw.csv`, `beta_dist_M_all_raw.csv` with per-mode splits `beta_dist_M_legacy_raw.csv` / `beta_dist_M_indep_raw.csv`, `size_adjusted_power.csv`, `ruwe_gate_all_raw.csv`, `gaia_quality_DR3.csv`, `confirm_independent_all_raw.csv`) and JSON summaries (`arm_delta_summary.json`, `beta_dist_summary.json`, `gate_lr_check.json`, `ruwe_clean_summary.json`, `confirm_summary.json`) are archived as ground truth in `results/`; replicate CSVs store the arm label `null` as text (read with `keep_default_na=False`).

Acknowledgments

The authors thank the maintainers of the NASA Exoplanet Archive, Gaia DR3, and the SWEET-Cat catalog for making this work possible. The authors also thank the independent adversarial audit passes of the analysis pipeline that identified methodological and implementation issues before the final results were reported.

Use of AI assisted technology. An AI assistant was used during preparation of this work for code review. All scientific design decisions, code execution, verification of reported numbers, and interpretations were carried out by the authors. The tool is not an author and bears no responsibility for the content.

Declarations

Competing Interests: The authors declare that they have no competing interests.

Funding: The authors received no specific funding for this work.

Authors' Contributions: Z.H.S. and M.M.A.A. designed the study, developed and audited the statistical pipeline, executed the archival queries and Monte Carlo calibration ensembles, verified all reported results, and wrote and approved the final manuscript.

Facilities: Exoplanet Archive, Gaia

Software: astropy (Astropy Collaboration 2013, 2018, 2022), numpy (Harris et al. 2020), scipy (Virtanen et al. 2020), gyrointerp (Angus et al. 2019)

References

Angus, R., T. D. Morton, D. Foreman-Mackey, et al. 2019. "Toward Precise Stellar Ages: Combining Isochrone Fitting with Empirical Gyrochronology." *The Astronomical Journal* 158: 173. https://doi.org/10.3847/1538-3881/ab3c53.

Astropy Collaboration. 2013. "Astropy: A community Python package for astronomy." *Astronomy & Astrophysics* 558: A33. https://doi.org/10.1051/0004-6361/201322068.

Astropy Collaboration. 2018. "The Astropy Project: Building an Open-science Project and Status of the v2.0 Core Package." *The Astronomical Journal* 156: 123. https://doi.org/10.3847/1538-3881/aabc4f.

Astropy Collaboration. 2022. "The Astropy Project: Sustaining and Growing a Community-oriented Open-source Project and the Latest Major Release (v5.0) of the Core Package." *The Astrophysical Journal* 935: 167. https://doi.org/10.3847/1538-4357/ac7c74.

Aumer, M., and J. J. Binney. 2009. "Kinematics and history of the solar neighbourhood revisited." *Monthly Notices of the Royal Astronomical Society* 397: 1286. https://doi.org/10.1111/j.1365-2966.2009.15053.x.

Bayer, Adrian E., and David N. Spergel. 2026. "Beyond χ^2: Ensemble null tests for reliable scientific discovery." *arXiv e-Prints*. https://arxiv.org/abs/2610.40167.

Berger, T. A., D. Huber, E. Gaidos, et al. 2020. "Revised Radii of Kepler Stars and Planets Using Gaia Data Release 2." *The Astronomical Journal* 160: 108. https://doi.org/10.3847/1538-3881/aba18a.

Bergsten, Galen J., Ilaria Pascucci, Gijs D. Mulders, et al. 2026. "Modeling the Impact of Unresolved Stellar Companions on Detection Sensitivity in Kepler's Small Planet Occurrence Rates." *arXiv e-Prints*. https://arxiv.org/abs/2601.16031.

Chen, D.-C., J.-W. Xie, J.-L. Zhou, et al. 2022. "Planets Across Space and Time (PAST) III: The Ages and Metallicities of Kepler Planet Host Stars." *The Astronomical Journal* 163: 249. https://doi.org/10.3847/1538-3881/ac5c4f.

Chen, Xunzhou, Tiancheng Sun, Yuxi Lu, Zixuan Lu, and Lifei Ye. 2026. “Insights into the Exoplanet Radius Valley from Host-Star Ages, Activity, Chemistry, and Birth Radii.” *The Astrophysical Journal* 1001: 176. https://doi.org/10.3847/1538-4357/ae5633.

Cloutier, Ryan, and Kristen Menou. 2020. “Evolution of the Radius Valley around Low-mass Stars from Kepler and K2.” *The Astronomical Journal* 159 (5): 211. https://doi.org/10.3847/1538-3881/ab8237.

David, T. J., G. Contardo, A. Sandoval, R. Angus, et al. 2021. “Evolution of the Exoplanet Size Distribution: Forming Large Super-Earths over Billions of Years.” *The Astronomical Journal* 161: 265. https://doi.org/10.3847/1538-3881/abf439.

Fulton, B. J., E. A. Petigura, A. W. Howard, et al. 2017. “The California-Kepler Survey. III. A Gap in the Radius Distribution of Small Planets.” *The Astronomical Journal* 154: 109. https://doi.org/10.3847/1538-3881/aa80eb.

Gaidos, E., A. Ali, A. L. Kraus, and J. F. Rowe. 2024. “The Radius Distribution of M dwarf-hosted Planets and its Evolution.” *Monthly Notices of the Royal Astronomical Society* 534: 3277. https://doi.org/10.1093/mnras/stae2207.

Gaidos, Eric, Zachary R. Claytor, Ryan Dungee, Aleezah Ali, and Gregory A. Feiden. 2023. “The TIME Table: rotation and ages of cool exoplanet host stars.” *Monthly Notices of the Royal Astronomical Society* 520: 5283. https://doi.org/10.1093/mnras/stad343.

Gillis, Erik Diego, Ryan Cloutier, and Emily K. Pass. 2026. “TESS Planet Occurrence Rates Reveal the Disappearance of the Radius Valley around Mid-to-late M Dwarfs.” *The Astronomical Journal* 171: 317. https://doi.org/10.3847/1538-3881/ae5810.

Ginzburg, S., H. E. Schlichting, and R. Sari. 2018. “Core-powered mass-loss and the radius distribution of small exoplanets.” *Monthly Notices of the Royal Astronomical Society* 476: 759. https://doi.org/10.1093/mnras/sty290.

Greco, Nino, A. C. Lanzafame, I. Pagano, G. Scandariato, V. Drago, and F. Spada. 2026. “Tracking the evolution of sub-Neptunes. II. The radius valley on gyrochronological ages.” *arXiv e-Prints*. https://arxiv.org/abs/2611.02099.

Gupta, A., and H. E. Schlichting. 2019. “Sculpting the valley in the radius distribution of small exoplanets as a by-product of planet formation: the core-powered mass-loss mechanism.” *Monthly Notices of the Royal Astronomical Society* 487: 24. https://doi.org/10.1093/mnras/stz1230.

Gupta, A., and H. E. Schlichting. 2020. “Signatures of the core-powered mass-loss mechanism in the exoplanet radius distribution.” *Monthly Notices of the Royal Astronomical Society* 493: 792. https://doi.org/10.1093/mnras/staa315.

Halsey, Lewis G., Douglas Curran-Everett, Sarah L. Vowler, and Gordon B. Drummond. 2015. “The fickle P value generates irreproducible results.” *Nature Methods* 12: 179–85. https://doi.org/10.1038/nmeth.3288.

Harris, Charles R., K. Jarrod Millman, Stefan J. van der Walt, et al. 2020. “Array programming with NumPy.” *Nature* 585: 357. https://doi.org/10.1038/s41586-020-2649-2.

Ho, Cynthia S. K., James G. Rogers, Vincent Van Eylen, James E. Owen, and Hilke E. Schlichting. 2024. “Shallower radius valley around low-mass hosts: evidence for icy planets, collisions, or high-energy radiation scatter.” *Monthly Notices of the Royal Astronomical Society* 531: 3698. https://doi.org/10.1093/mnras/stae1376.

Ho, Cynthia S. K., and Vincent Van Eylen. 2023. “A deep radius valley revealed by Kepler short cadence observations.” *Monthly Notices of the Royal Astronomical Society* 519 (3): 4056–73. https://doi.org/10.1093/mnras/stac3802.

Junk, Thomas R., and Louis Lyons. 2020. “Reproducibility and Replication of Experimental Particle Physics Results.” *Harvard Data Science Review*, ahead of print. https://doi.org/10.1162/99608f92.250f995b.

Kamulali, J., V. Adibekyan, B. Nsamba, et al. 2026. “Revisiting the exoplanet radius valley with host stars from SWEET-Cat.” *Astronomy & Astrophysics* 707: A41. https://doi.org/10.1051/0004-6361/202557554.

Lee, E. J., and E. Chiang. 2016. “Breeding Super-Earths and Birthing Super-puffs in Transitional Disks.” *The Astrophysical Journal* 817: 90. https://doi.org/10.3847/0004-637X/817/2/90.

Lopez, E. D., and J. J. Fortney. 2013. “The Role of Core Mass in Controlling Evaporation: The Kepler Radius Distribution and the Kepler-36 Density Dichotomy.” *The Astrophysical Journal* 776: 2. https://doi.org/10.1088/0004-637X/776/1/2.

Lopez, E. D., and K. Rice. 2018. “Predictions for the period dependence of the transition between rocky super-Earths and gaseous sub-Neptunes and implications for TESS.” *Monthly Notices of the Royal Astronomical Society* 479: 5303. https://doi.org/10.1093/mnras/sty1707.

Lu, Yuxi, and Marc H. Pinsonneault. 2026. “Ensemble Kinematic Ages for 1.5 Million LAMOST Stars.” *arXiv e-Prints*. https://arxiv.org/abs/2607.22309.

Luque, Rafael, and Enric Pallé. 2022. “Density, not radius, separates rocky and water-rich small planets orbiting M dwarf stars.” *Science* 377 (6611): 1211–14. https://doi.org/10.1126/science.abl7164.

Mallik, Ananya. 2026. “Orbit-Hausdorff consistency of maximum likelihood.” *arXiv e-Prints*. https://arxiv.org/abs/2509.12206.

McQuillan, A., T. Mazeh, and S. Aigrain. 2014. “Rotation Periods of 34,030 Kepler Main-sequence Stars.” *The Astrophysical Journal Supplement Series* 211: 24. https://doi.org/10.1088/0067-0049/211/2/24.

Ndaoud, Mohamed. 2022. “Sharp optimal recovery in the two-component Gaussian mixture model.” *Annals of Statistics* 50: 2096–126. https://doi.org/10.1214/22-AOS2178.

Owen, J. E., and Y. Wu. 2017. “The Evaporation Valley in the Kepler Planets.” *The Astrophysical Journal* 847: 29. https://doi.org/10.3847/1538-4357/aa890a.

Parashivamurthy, Harshitha M., and Gijs D. Mulders. 2025. “Radius valley scaling among low-mass stars with TESS.” *Astronomy & Astrophysics* 703: A8. https://doi.org/10.1051/0004-6361/202554006.

Petigura, Erik A., James G. Rogers, Howard Isaacson, et al. 2022. “The California-Kepler Survey. X. The Radius Gap as a Function of Stellar Mass, Metallicity, and Age.” *The Astronomical Journal* 163 (4): 179. https://doi.org/10.3847/1538-3881/ac51e3.

Rauer, Heike, Conny Aerts, Juan Cabrera, et al. 2025. “The PLATO mission.” *Experimental Astronomy* 59: 26. https://doi.org/10.1007/s10686-025-09985-9.

Rogers, James G. 2025. “On the road to the radius valley: distinguishing gas dwarfs from water worlds with young transiting exoplanets.” *Monthly Notices of the Royal Astronomical Society* 539: 2230–41. https://arxiv.org/abs/2503.17364.

Sandoval, Angeli, Gabriella Contardo, and Trevor J. David. 2021. “The Influence of Age on the Relative Frequency of Super-Earths and Sub-Neptunes.” *The Astrophysical Journal* 911: 117. https://doi.org/10.3847/1538-4357/abea9e.

Sayeed, Maryum, Ruth Angus, Travis A. Berger, et al. 2025. “Exoplanet Occurrence Rate with Age for FGK Stars in Kepler.” *The Astronomical Journal* 169: 112. https://doi.org/10.3847/1538-3881/ada8a1.

Van Eylen, V., C. Agentoft, M. S. Lundkvist, et al. 2018. “An asteroseismic view of the radius valley: stripped cores, not born rocky.” *Monthly Notices of the Royal Astronomical Society* 479: 4786. https://doi.org/10.1093/mnras/sty1783.

Venturini, J., M. P. Ronco, O. M. Guilera, J. Haldemann, C. Mordasini, and M. Miller Bertolami. 2024. “A fading radius valley towards M dwarfs, a persistent density valley across stellar types.” *Astronomy & Astrophysics* 686: L9. https://doi.org/10.1051/0004-6361/202349088.

Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, et al. 2020. “SciPy 1.0: fundamental algorithms for scientific computing in Python.” *Nature Methods* 17: 261. https://doi.org/10.1038/s41592-019-0686-2.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryData1FGKgaterawreplicates1000perarm.csv
- SupplementaryData3powergridrawreplicates500percell.csv
- SUPPLEMENTARYINDEX.md