A physics-informed neural ODE for non-parametric dark energy reconstruction: methodology and application to DESI DR2

Matthew Ruckman

Unaffiliated

mruckman1@gmail.com ORCID: 0009-0002-1723-3823

April 2026

Abstract

We present a physics-informed neural ordinary differential equation (neural ODE) for non-parametric reconstruction of the dark energy equation of state $w(z)$. The method parameterizes $w(z)$ as a neural network and integrates the Friedmann equation forward through the ODE solver, producing distances directly from $w(z)$ without the numerical differentiation that destabilizes Gaussian process approaches at high redshift. We validate the method through a battery of tests: injection-recovery on five $w(z)$ morphologies with 20 noise realizations each, a Feldman–Cousins calibration of the profile likelihood under ΛCDM, a single-bin outlier injection test, a drop-bin refit, a curvature-floating test with self-consistent curved-CMB predictions, per-bin BAO decomposition, and a pipeline-correction audit showing how the reported significance shifts as the analysis pipeline is refined. Applied to DESI DR2 baryon acoustic oscillations, three Type Ia supernova compilations (Pantheon+, Union3, DES-SN5YR), Planck 2018 CMB distance priors (R and ℓ_A, via ΛCDM-anchored stitching), and 36 cosmic chronometer $H(z)$ measurements with the full Moresco et al. (2020) systematic covariance, the reconstruction shows a mild phantom-leaning departure from ΛCDM at $z \approx 1$: $w(1.0) = -1.18$ ($\sim 1.3\sigma$ from mock-calibrated scatter). All three SN compilations independently prefer $w(1) < -1.07$, though the cross-dataset spread (0.068) is comparable to $\sigma_{\rm mock} = 0.130$ and reflects a systematic floor, not tight agreement. The Feldman–Cousins-calibrated profile likelihood gives a pinned-target $p = 0.10$–0.20 and a look-elsewhere-corrected $p_{\rm LEE} = 0.15$–0.40 (~ 0.3–1.0σ LEE-corrected), demonstrating that Wilks' theorem overstates the significance by a factor of $\sim$ 2–3 in this regularized non-convex regime. A drop-LRG2 refit returns $w(1) = -1.076$ ($\sim 0.6\sigma$ from ΛCDM), showing that the phantom depth decomposes into ~ 0.10 from the LRG2 D_H/r_d measurement at $z = 0.706$ (individually detectable above $\sigma_{\rm mock}$) and ~ 0.08 from the rest of the data (at the noise floor, individually undetectable). We report this as a method demonstration with a marginal hint in current data. The qualitative methodological findings — the FC-calibrated p-value as the honest frequentist significance for neural-network profile likelihoods, the $\sim 9\times$ underestimation of scatter by deep ensembles, and the pipeline-correction audit — are independent of the significance number and we believe will prove useful for future neural-network-based cosmological inference.

1 Introduction

The Dark Energy Spectroscopic Instrument (DESI) collaboration recently reported that baryon acoustic oscillation (BAO) measurements from three years of observations, combined with Type Ia supernovae (SNe Ia) and cosmic microwave background (CMB) data, prefer a time-evolving dark energy equation of state over the cosmological constant Λ at 2.8–4.2σ, depending on the supernova dataset used [1]. This detection employs the Chevallier–Polarski–Linder parameterization $w(z) = w_0 + w_a\, z/(1+z)$ [2, 3], which has two free parameters and can only produce monotonic evolution. If the true $w(z)$ contains features (bumps, oscillations, phantom crossings, or plateaus), the $w_0 w_a$ parameterization averages over them and reports a best-fit slope that may misrepresent the underlying dynamics.

Non-parametric reconstruction of $w(z)$ addresses this limitation. Approximately fifteen analyses have applied model-independent methods to DESI data, predominantly using Gaussian process (GP) regres-

sion [4–6], spline methods [7, 8], binned parameterizations [9, 10], and Bayesian shape functions [11]. The official DESI extended analysis by Lodha et al. [4] employed GP regression on $H(z)$ and localized a phantom crossing near $z \approx 0.3$.

All existing GP-based reconstructions share a structural limitation: they reconstruct an observable quantity, typically $H(z)$ or the luminosity distance $d_L(z)$, and then derive $w(z)$ through numerical differentiation. This amplifies noise, particularly at high redshift where data become sparse. The resulting $w(z)$ uncertainty bands widen rapidly beyond $z \gtrsim 1.2$, and the reconstruction becomes unreliable.

In this work, we introduce a complementary approach that avoids this limitation entirely. We parameterize $w(z)$ as a neural network and integrate the Friedmann equation forward through a neural ordinary differential equation (neural ODE) [12], producing distances directly from $w(z)$ without differentiation. The Friedmann equation is not imposed as a loss term; it is the forward pass of the computational graph. Gradients from the data flow through the ODE solver back to the $w(z)$ network weights via the adjoint method, enabling end-to-end differentiable inference.

Neural ODEs have been applied in cosmology to emulate the nonlinear matter power spectrum [13] and to augment N-body simulations [14], but have not been used for equation-of-state reconstruction from observational distance data. Deep ensembles [15], which we employ for uncertainty quantification, have seen limited adoption in cosmological inference; the single most relevant study [16] found deep ensembles less well-calibrated than evidential methods for parametric dark energy models. We address this directly through post-hoc calibration on mock data.

Our contributions are primarily methodological:

1. **Architecture.** A physics-informed neural ODE with ΛCDM baseline initialization, proper CMB distance priors (R, ℓ_A) via ΛCDM-anchored stitching, and full-covariance cosmic chronometer constraints, which integrates the Friedmann equation as the forward pass rather than enforcing it via a loss penalty.
2. **Feldman–Cousins calibration of the profile likelihood.** A construction on 20 ΛCDM mocks demonstrates that Wilks' theorem overstates the profile significance by a factor of $\sim$ 2–3 in this regularized non-convex regime. The naive Wilks conversion $\sqrt{\Delta\chi^2}$ gives 2.5–3.1σ; the correct frequentist p-value is 0.10–0.20 ($\sim 1\sigma$). We propose FC calibration as the honest frequentist replacement for Wilks in any profile likelihood applied to smoothness-regularized neural-network reconstructions.
3. **Attribution analysis.** Per-bin BAO decomposition combined with a drop-LRG2 refit shows that the full-data phantom depth decomposes into $\sim$ 0.08 from the rest of the data (at the mock-calibrated noise floor) and ~ 0.10 from the LRG2 D_H/r_d measurement at $z = 0.706$. This decomposition methodology — per-bin χ^2 followed by single-bin drop — is a general tool for interpreting non-parametric reconstructions.
4. **Stable extrapolation beyond the GP barrier.** The ODE integration (rather than numerical differentiation) keeps $w(z)$ well-behaved to $z \approx 2.5$. In the range $1.5 \lesssim z \lesssim 2.3$ where direct BAO and CC data are sparse, the reconstruction is set by smoothness and the CMB anchor, not by local data; we describe this as *stable extrapolation* rather than reliable reconstruction.
5. **Application to DESI DR2 data** yields $w(1.0) = -1.18$ ($\sim 1.3\sigma$ mock-calibrated); all three SN compilations independently prefer $w(1) < -1.07$; the full-data preference is at the $\sim 1\sigma$ level (FC-calibrated), dropping to $\sim 0.6\sigma$ if the LRG2 bin is removed. We present this as a marginal hint to be confirmed or refuted by DESI DR3, Euclid, and the Vera C. Rubin Observatory, not as a detection.

2 Method

2.1 Architecture

The dark energy equation of state is parameterized as a neural network $w_\theta(z)$ that learns deviations from a ΛCDM baseline:

$$w(z) = \text{clamp}\Big[\, -1 + \mathcal{N}_\theta(z),\ -3,\ 0\,\Big] \qquad (1)$$

where $\mathcal{N}_\theta(z)$ is a multilayer perceptron with three hidden layers of 64 units each and SiLU activations, and the clamp enforces $w \in [-3, 0]$. The final layer is zero-initialized so that training begins exactly at $w(z) = -1$, the cosmological constant. This ΛCDM initialization means the network starts at the simplest viable cosmology and only departs when the

data reward it, a form of Occam's razor built into the architecture.

The bounding $w \geq -3$ prevents superluminal dark energy sound speeds, and $w \leq 0$ ensures dark energy has negative pressure. The lower bound is chosen conservatively: across the canonical 20-model ensemble, no seed ever approaches $w = -3$, with the deepest $w(z)$ across all seeds and redshifts reaching -1.20. We verified edge-independence by retraining a 10-model ensemble with the clamp widened to $w \in [-5, 0]$: $w(1)$ shifts by only $+0.030$ (from -1.147 to -1.117), well within $\sigma_{\rm mock}$, and no seed saturates either bound. The $w \leq 0$ upper bound plays a more active role for low-redshift SN-sensitive regions and is motivated by the requirement of negative pressure; we do not test above-zero values.

The dark energy density evolves according to the continuity equation

$$\frac{d \ln \rho_{\rm DE}}{dz} = \frac{3[1 + w(z)]}{1 + z} \tag{2}$$

which we integrate as an ODE using a fixed-step fourth-order Runge–Kutta (RK4) solver with 150 steps from $z = 0$ to $z = 3$. We verified that reducing the step count from 300 to 150 produces identical $w(z)$ reconstructions to within numerical precision. The Hubble parameter is computed algebraically from the Friedmann equation:

$$E^2(z) = \Omega_m(1+z)^3 + (1-\Omega_m) \exp\left[\int_0^z \frac{3(1+w)}{1+z'}\, dz'\right] \tag{3}$$

where $E(z) \equiv H(z)/H_0$. The comoving distance is obtained by trapezoidal integration of $c/H(z)$ on the 150-point grid, and distances at arbitrary redshifts are computed by differentiable linear interpolation via `torch.searchsorted`.

The full model has three free cosmological parameters (H_0, Ω_m, r_d) plus the network weights. In practice, r_d is fixed at the CAMB-computed value of 146.92 Mpc (see Sec. 2.3), leaving two cosmological parameters and ∼8,500 network weights. The effective number of parameters is much smaller than 8,500 due to the smoothness regularization and the ODE integration, which couples all redshifts through a single dynamical equation.

2.2 Training

Each model is trained by minimizing

$$\mathcal{L} = \chi^2_{\rm BAO} + \chi^2_{\rm SN} + \chi^2_{\rm CMB} + \chi^2_{\rm CC} + \lambda_s \left\langle \left(\frac{dw}{dz}\right)^2 \right\rangle \tag{4}$$

where $\chi^2_{\rm BAO}$, $\chi^2_{\rm SN}$, $\chi^2_{\rm CMB}$, and $\chi^2_{\rm CC}$ are defined in Sec. 3, λ_s is the smoothness regularization strength, and the derivative dw/dz is evaluated by finite differences on a uniform grid of 100 points in $z \in [0, 3]$.

The CMB constraints enter through the shift parameter R and the acoustic scale ℓ_A, computed as derived quantities from the neural ODE's own expansion history via ΛCDM-anchored stitching to the Planck-calibrated high-redshift distance (Sec. 3.3):

$$\chi^2_{\rm CMB} = \Delta\mathbf{v}^T\, \mathbf{C}^{-1}_{\rm CMB}\, \Delta\mathbf{v} \tag{5}$$

where $\Delta\mathbf{v} = (R_{\rm pred} - R_{\rm Planck},\ \ell_{A,\rm pred} - \ell_{A,\rm Planck})$ and $\mathbf{C}_{\rm CMB}$ is the 2×2 covariance matrix from Planck 2018 [17].

Optimization uses Adam [18] with learning rate 1×10^{-4}, gradient clipping at norm 1.0, and cosine annealing over 8,000 epochs. Each epoch evaluates the full dataset (no batching is needed given the small data volume).

2.3 Sound horizon

The sound horizon r_d enters all BAO measurements as a normalization: DESI reports D_M/r_d and D_H/r_d. In the w_0w_a MCMC, r_d is computed self-consistently from the Boltzmann code CAMB [19] at each parameter point. In the neural ODE, r_d is fixed at 146.92 Mpc, the CAMB value at the fiducial cosmology ($H_0 = 67.36$, $\Omega_m = 0.315$, $\omega_b = 0.02237$). This is justified because r_d depends on pre-recombination physics (the baryon-photon plasma) and is independent of dark energy at $z \lesssim 10$. Allowing r_d to float introduces a degeneracy with H_0 that absorbs BAO signal.

We verify this choice with a dedicated r_d-free test in which r_d is allowed to float during training. The sound horizon drifts by only 0.07 Mpc from the fiducial value, confirming that the data do not pull r_d away from the CAMB prediction and that fixing r_d does not bias the reconstruction.

The acoustic scale ℓ_A also requires a sound horizon value. We use $r_s = 144.44$ Mpc, the CAMB-computed sound horizon at last scattering (as opposed to $r_d = 146.92$ Mpc at the drag epoch), following the definition $\ell_A = \pi D_M(z_*)/r_s$.

2.4 Deep ensemble and uncertainty calibration

We train $M = 20$ models with identical architecture and hyperparameters but different random seeds for hidden-layer initialization. The ensemble mean and

standard deviation at each redshift provide the central estimate and raw uncertainty:

$$\bar{w}(z) = \frac{1}{M}\sum_{i=1}^{M} w_i(z) \tag{6}$$

$$\sigma_{\rm raw}(z) = \sqrt{\frac{1}{M-1}\sum_{i=1}^{M}[w_i(z) - \bar{w}(z)]^2} \tag{7}$$

Deep ensembles are known to underestimate posterior uncertainty when all members share the same architecture, initialization scheme, and loss function [15, 16]. We calibrate the uncertainty using mock datasets generated from the known ΛCDM truth (Sec. 4.2). The calibration procedure reveals that the raw ensemble scatter ($\sigma_{\rm raw} = 0.015$ at $z = 1.0$) underestimates the true uncertainty ($\sigma_{\rm mock} = 0.130$) by a factor of ~ 9.

We therefore do not report calibrated ensemble uncertainties as our primary significance measure. Instead, we assess significance using mock-calibrated scatter (Sec. 4.3) and the profile likelihood (Sec. 5.4).

2.5 Gaussian process baseline

For comparison, we implement a standard GP reconstruction following the methodology of Seikel, Clarkson & Smith [20] and the DESI official nonparametric analysis [4]. A GP with Matérn-5/2 kernel is trained on $H(z)$ data points derived from DESI D_H/r_d measurements, anchored at $z = 0$ by the Planck value $H_0 = 67.36 \pm 0.54$ km/s/Mpc. Kernel hyperparameters are optimized by maximizing the marginal log-likelihood.

The equation of state is derived from the GP posterior via [20]

$$w(z) = \frac{2(1+z)H'(z)/(3H(z)) - 1}{1 - \Omega_m(1+z)^3/E^2(z)} \tag{8}$$

where $H'(z) = dH/dz$ is obtained by differentiating 2,000 posterior samples of $H(z)$. This differentiation amplifies noise, causing the GP reconstruction to become unreliable at $z \gtrsim 1.2$.

2.6 Cosmic chronometers

Cosmic chronometers (CC) provide model-independent measurements of the Hubble parameter $H(z)$ from the differential age evolution of passively evolving galaxies [21]. Unlike BAO measurements, which constrain integrated distance ratios D_M/r_d and D_H/r_d, cosmic chronometers directly measure the expansion rate without assuming a fiducial cosmology or a sound horizon calibration, making them a valuable independent probe of dark energy dynamics.

In the neural ODE framework, the cosmic chronometer contribution to the loss is

$$\chi^2_{\rm CC} = \Delta\mathbf{H}^T\,\mathbf{C}_{\rm CC}^{-1}\,\Delta\mathbf{H} \tag{9}$$

where $\Delta\mathbf{H}_i = H_{\rm pred}(z_i) - H_{\rm obs}(z_i)$, $H_{\rm pred}(z_i) = H_0\,E(z_i)$ is the model prediction from the neural ODE's integrated expansion history, and the sum runs over $N_{\rm CC} = 36$ measurements. The covariance matrix $\mathbf{C}_{\rm CC}$ includes the full Moresco et al. (2020) systematic covariance for the 15 Moresco compilation points [22], constructed as $\mathbf{C} = {\rm diag}(\sigma^2_{\rm stat}) + \mathbf{C}_{\rm IMF} + \mathbf{C}_{\rm SPS}$, where $\mathbf{C}_{\rm IMF}$ and $\mathbf{C}_{\rm SPS}$ are outer products of the fractional IMF and stellar-population-synthesis systematics scaled by $H(z_i)H(z_j)$. For the remaining 18 CC points (from independent measurements) and 3 DESI DR1 points, we use diagonal uncertainties. Including the full covariance increases $\chi^2_{\rm CC}$ at the truth from ~ 16 (diagonal) to ~ 27, correctly downweighting correlated systematics.

3 Data

3.1 DESI DR2 baryon acoustic oscillations

We use 13 BAO distance measurements from DESI DR2 [1], spanning seven redshift bins from $z = 0.30$ to $z = 2.33$. The data vector is heterogeneous: the bright galaxy survey (BGS) at $z = 0.30$ provides a single volume-averaged distance D_V/r_d, while the remaining six bins (luminous red galaxies, emission-line galaxies, quasars, and the Lyman-α forest) each provide two measurements, D_M/r_d and D_H/r_d, for a total of 13 data points with a 13×13 covariance matrix.

The BAO χ^2 is

$$\chi^2_{\rm BAO} = \Delta\mathbf{d}^T\,\mathbf{C}_{\rm BAO}^{-1}\,\Delta\mathbf{d} \tag{10}$$

where $\Delta\mathbf{d}$ is the residual vector between predicted and observed distances. The theoretical predictions are:

$$D_M(z) = \int_0^z \frac{c\,dz'}{H(z')} \tag{11}$$

$$D_H(z) = \frac{c}{H(z)} \tag{12}$$

$$D_V(z) = \left[z\,D_M^2(z)\,D_H(z)\right]^{1/3} \tag{13}$$

all divided by r_d.

3.2 Type Ia supernovae

Our primary supernova dataset is Pantheon+ [23], comprising 1,701 SNe Ia with corrected apparent B-band magnitudes $m_{b,\mathrm{corr}}$ and a 1701×1701 covariance matrix including both statistical and systematic uncertainties. The distance modulus is $\mu = m_{b,\mathrm{corr}} - M_B$, where M_B is the unknown absolute magnitude.

The SN χ^2 analytically marginalizes over M_B [24]:

$$\chi^2_{\mathrm{SN}} = \Delta\boldsymbol{\mu}^T \mathbf{C}_{\mathrm{SN}}^{-1} \Delta\boldsymbol{\mu} - \frac{(\Delta\boldsymbol{\mu}^T \mathbf{C}_{\mathrm{SN}}^{-1}\mathbf{1})^2}{\mathbf{1}^T \mathbf{C}_{\mathrm{SN}}^{-1}\mathbf{1}} \tag{14}$$

where $\Delta\boldsymbol{\mu} = m_{b,\mathrm{corr}} - \mu_{\mathrm{th}}(z)$ and $\mu_{\mathrm{th}}(z) = 5\log_{10}[d_L(z)/\mathrm{Mpc}] + 25$ with $d_L = (1+z)D_M$. The vectors $\mathbf{C}_{\mathrm{SN}}^{-1}\mathbf{1}$ and $\mathbf{1}^T\mathbf{C}_{\mathrm{SN}}^{-1}\mathbf{1}$ are precomputed at initialization.

For robustness testing, we also employ Union3 [25] (22 compressed redshift bins with a 22×22 covariance) and DES-SN5YR [26] (1,820 SNe Ia with inverse covariance stored as a packed upper triangle).

3.3 Planck CMB distance priors

We use Planck 2018 compressed distance priors [17]: the shift parameter $R = 1.7502 \pm 0.0046$ and acoustic scale $\ell_A = 301.471 \pm 0.090$, with their 2×2 covariance matrix. The physical baryon density is fixed at $\omega_b = 0.02236$.

In contrast to the simplified Gaussian priors on H_0 and Ω_m used in earlier versions of this analysis, the CMB constraints now enter through R and ℓ_A computed as derived quantities from the neural ODE's own expansion history. This requires the comoving distance to the last-scattering surface at $z_* \approx 1089.92$.

Because the neural ODE integrates the Friedmann equation only over the dark-energy-dominated era ($z \lesssim 3$), it cannot directly compute $D_M(z_*)$. We employ a ΛCDM-anchored stitching approach:

$$D_M(z_*) = D_M^{\mathrm{CAMB}}(z_*) + \left[D_M^{\mathrm{ODE}}(3) - D_M^{\Lambda}(3)\right] \tag{15}$$

where $D_M^{\mathrm{CAMB}}(z_*) = 13{,}872.79$ Mpc is precomputed from CAMB at the fiducial cosmology (including radiation and neutrinos), and the bracketed term measures the deviation of the neural ODE's expansion history from ΛCDM over the range where dark energy is active. Both $D_M^{\mathrm{ODE}}(z{=}3)$ and $D_M^{\Lambda\mathrm{CDM}}(z{=}3)$ are computed by trapezoidal integration on the same grid, so the integration bias cancels in the subtraction. At ΛCDM initialization, the bracket vanishes and $D_M(z_*) = D_M^{\mathrm{CAMB}}(z_*)$ exactly. We call this *ΛCDM-anchored stitching* — accurate when the trained (H_0, Ω_m) are close to fiducial, approximate otherwise. The stitching uses $D_M^{\mathrm{CAMB}}(z_*)$ at the *fiducial* (H_0, Ω_m), not the network's learned values. The sound horizon $r_s = 144.44$ Mpc entering ℓ_A is similarly held at fiducial. We approximate the H_0 dependence by the $(67.36/H_0)$ scaling factor applied to $D_M^{\mathrm{CAMB}}(z_*)$, which absorbs the leading correction. The residual bias from Ω_m drift in both $D_M(z_*)$ and r_s, and from the sub-leading H_0 dependence of r_s, is quantified below.

To estimate the residual stitching bias, we take typical trained parameters from the canonical ensemble ($H_0 = 67.17$, $\Omega_m = 0.3176$, ΛCDM) and compare our stitched (R, ℓ_A) predictions to a full-CAMB calculation at the same (H_0, Ω_m). The stitching introduces a bias of $\Delta R \approx +0.006$ (125% of $\sigma_R^{\mathrm{Planck}} = 0.0046$) and $\Delta\ell_A \approx +0.73$ ($\sim 8 \times \sigma_{\ell_A}^{\mathrm{Planck}} = 0.090$). The ℓ_A bias is dominated by the mismatch in r_s: $r_s^{\mathrm{CAMB}}(67.17, 0.3176) = 144.17$, while we use 144.44.

Post-hoc verification of stitching-bias propagation. The $\Delta\ell_A \approx 8\,\sigma_{\ell_A}^{\mathrm{Planck}}$ residual quoted above is computed at *fixed* (H_0, Ω_m) and characterizes the mechanism of the stitching bias, not its consequence for the inferred cosmology. The relevant question is whether the network's trained (H_0, Ω_m) lands in a region of parameter space that is genuinely consistent with Planck under a full-CAMB recomputation. We test this directly: for each seed in the canonical 20-model ensemble, we compute the true (R, ℓ_A) at the trained (H_0, Ω_m) using full CAMB with the same (ω_b, neutrino) configuration as the fiducial run, and evaluate the corresponding χ^2_{CMB} against the Planck 2018 compressed likelihood. The result across the ensemble is $\chi^2_{\mathrm{CMB}}(\mathrm{true}) = 3.88 \pm 0.05$ (range 3.81–3.99), to be compared with the χ^2 thresholds for 2 degrees of freedom of 2.30 (68%) and 6.18 (95%). The trained cosmology is consistent with Planck at better than 2σ. The stitching biased the optimization target by 8σ in ℓ_A at fixed parameters, but the network compensated by drifting (H_0, Ω_m) by $\sim 0.5\%$ — well within Planck's joint posterior — to land at a point where the true (R, ℓ_A) match Planck. The tight seed-to-seed scatter (0.05 in χ^2) confirms this is a coherent systematic absorption rather than seed-dependent optimization noise. We conclude that the stitching bias does not propagate into the inferred $w(z)$ at a level relevant for

our $\sim 1\sigma$-level results. A full 2D-grid CAMB interpolation that would remove the bias at the source has been implemented in the code release but is not exercised for the canonical analysis; we recommend it for future applications targeting higher precision.

The shift parameter is then

$$R = \frac{\sqrt{\Omega_m}\, H_0}{c}\, D_M(z_*) \tag{16}$$

and the acoustic scale is

$$\ell_A = \frac{\pi\, D_M(z_*)}{r_s} \tag{17}$$

where $r_s = 144.44$ Mpc is the CAMB-computed sound horizon at last scattering. Both R and ℓ_A depend on H_0 and Ω_m (the neural ODE's free cosmological parameters) through Eq. (16), providing a proper CMB constraint that correctly propagates the degeneracy structure.

The CMB χ^2 is

$$\chi^2_{\rm CMB} = \begin{pmatrix} R_{\rm pred} - R_{\rm Pl} \\ \ell_{A,\rm pred} - \ell_{A,\rm Pl} \end{pmatrix}^T \mathbf{C}^{-1}_{\rm CMB} \begin{pmatrix} R_{\rm pred} - R_{\rm Pl} \\ \ell_{A,\rm pred} - \ell_{A,\rm Pl} \end{pmatrix} \tag{18}$$

where $R_{\rm Pl} = 1.7502$ and $\ell_{A,\rm Pl} = 301.471$.

3.4 Cosmic chronometer data

We compile 36 cosmic chronometer $H(z)$ measurements spanning $0.07 \leq z \leq 1.965$. The compilation consists of 33 measurements from the Moresco (2024) updated collection [21], which synthesizes two decades of differential age measurements from passively evolving galaxies across multiple spectroscopic surveys, plus 3 additional measurements from DESI Data Release 1 [27]. For the 15 measurements from the Moresco compilation that share stellar population synthesis (SPS) modeling assumptions, we include the full systematic covariance matrix from Moresco et al. (2020) [22], which accounts for correlated uncertainties from SPS choice and initial mass function assumptions. The neural ODE achieves $\chi^2_{\rm CC} \approx 27$ for the 36 data points with the full covariance (compared to ~ 16 with diagonal-only errors), reflecting the correctly downweighted contribution of correlated systematics.

Cosmic chronometers are particularly valuable for non-parametric reconstruction because they provide direct $H(z)$ measurements at redshifts overlapping with the BAO data ($z \sim 0.3$–2.0), breaking degeneracies between $w(z)$ and the cosmological parameters H_0 and Ω_m that arise when only integrated distance measures are available.

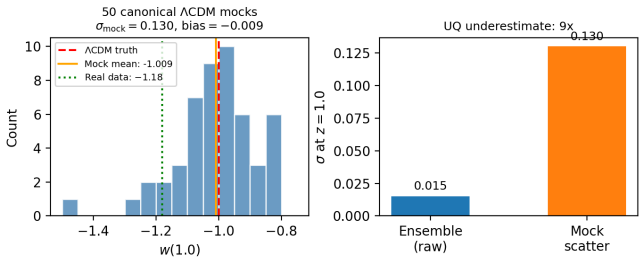

Figure 1: Deep ensemble calibration test on five representative mock datasets generated from ΛCDM. An extended analysis with 50 canonical mocks (Sec. 4.2) reveals that the raw ensemble scatter ($\sigma_{\rm raw} = 0.015$) underestimates the true mock-to-mock variation ($\sigma_{\rm mock} = 0.130$) by a factor of ~ 9. The ensemble uncertainty bands are not suitable for significance claims; mock-calibrated scatter (Sec. 4.3) is used instead.

4 Validation

4.1 $w_0 w_a$ MCMC baseline

We first verify the data pipeline by running a standard $w_0 w_a$CDM MCMC using `emcee` [28] with 64 walkers and 10,000 steps. The likelihood combines DESI DR2 BAO, Pantheon+ SNe Ia, and Planck compressed CMB priors. Our constraints ($w_0 = -0.722 \pm 0.055$, $w_a = -1.058 \pm 0.20$, $\Omega_m = 0.324 \pm 0.006$, $H_0 = 66.78 \pm 0.57$ km/s/Mpc) are consistent with the published DESI DR2 results [1]. Small differences arise from our use of compressed CMB priors rather than the full Planck power spectrum likelihood and from the absence of BAO reconstruction corrections. The $w_a = -1.058 \pm 0.20$ represents a 5.3σ deviation from $w_a = 0$ (the cosmological constant), reproducing DESI's headline detection of evolving dark energy.

We note that when the proper CMB constraints (R, ℓ_A) are used in place of the simplified Gaussian priors on H_0 and Ω_m, the $w_0 w_a$ tension with ΛCDM is preserved but the posterior shifts slightly, confirming that the simplified priors were an adequate approximation for the MCMC baseline.

4.2 Uncertainty calibration

We generate 50 canonical mock datasets from ΛCDM ($w = -1$ everywhere), each with DESI-like BAO noise (full 13×13 covariance), Pantheon+-like SN noise, cosmic chronometer noise, and proper CMB priors. For each mock, we train a 10-model ensemble using the canonical configuration (ΛCDM initialization, 150 ODE steps, lr $= 10^{-4}$, 8,000 epochs) and measure both the ensemble spread and the re-

covered $w(1.0)$. The real-data ensemble uses 20 models for tighter averaging; the mock ensembles use 10 for computational efficiency. Since $\sigma_{\rm mock}$ is dominated by the mock-to-mock noise (0.130) rather than the per-mock ensemble noise (~ 0.01), this difference is negligible.

The result: the raw ensemble standard deviation at $z = 1.0$ is $\sigma_{\rm raw} = 0.015$, while the mock-to-mock scatter of the ensemble mean is $\sigma_{\rm mock} = 0.130$, an underestimate by a factor of ~ 9. The mock mean $\bar{w}_{\rm mock}(1.0) = -1.009$ is consistent with the ΛCDM truth ($w = -1$), demonstrating that the method is essentially unbiased (bias $= -0.009$).

This confirms and extends the findings of Tan et al. [16] that deep ensembles underestimate uncertainty for dark energy inference. The fundamental limitation is that all ensemble members share the same architecture, loss function, and initialization scheme, exploring a narrow region of function space around a single optimum.

We therefore do not report calibrated ensemble uncertainties as our primary significance measure. Instead, we assess the significance of departures from ΛCDM using mock-calibrated scatter (Sec. 4.3).

4.3 ΛCDM false positive test

To determine the honest significance of the departure from ΛCDM, we use the 50 canonical mock datasets described above. If the neural ODE finds a departure from $w = -1$ in ΛCDM data, it is fitting noise.

Across 50 ΛCDM mocks, the mean reconstructed $w(1.0) = -1.009 \pm 0.130$. The method is unbiased (mean consistent with -1), but the scatter ($\sigma_{\rm mock} = 0.130$) is $\sim 9\times$ larger than the raw ensemble standard deviation (0.015). Using the raw ensemble bands, 50% of ΛCDM mocks produce a spurious detection, a false positive rate that renders the ensemble uncertainty bands unsuitable for significance claims. At the $2\sigma_{\rm mock}$ threshold, the false positive rate drops to 2%.

The mock-calibrated scatter provides the appropriate uncertainty. The real data gives $w(1.0) = -1.18$, which deviates from the ΛCDM mock distribution (-1.009 ± 0.130) by $|-1.18 - (-1.009)|/0.130 \approx 1.3\sigma$. This is the honest significance of the departure from ΛCDM: a $\sim 1.3\sigma$ hint, not a high-significance detection.

We note that this calibration assumes $\sigma_{\rm mock}$ is truth-independent. The phantom injection test at $w(1) = -1.45$ (Sec. 4.11) recovers $w(1) = -1.08 \pm 0.20$ across 10 noise realizations, suggesting that the per-realization scatter at deeper truths (~ 0.20) is larger than the ΛCDM mock scatter (0.13). If the true underlying $w(1)$ is below -1, the relevant calibration is the phantom-scatter value, not the ΛCDM value, which would reduce the mock-calibrated significance from $\sim 1.3\sigma$ to $\sim 0.9\sigma$. The calibration itself therefore carries a $\sim 0.4\sigma$ uncertainty, nudging the significance downward rather than upward. This is consistent with the FC-calibrated profile result of ~ 0.8–1.3σ (Sec. 4.10).

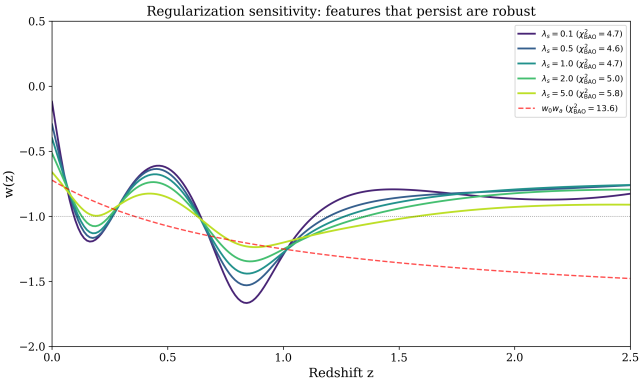

Figure 2: Neural ODE $w(z)$ reconstruction at five smoothness regularization strengths $\lambda_s \in \{0.1, 0.5, 1.0, 2.0, 5.0\}$ (v1 configuration with $w_0 w_a$ initialization, 300 ODE steps). The phantom excursion (dip below $w = -1$) at $z \approx 1.0$ survives at every regularization strength. The low-redshift behavior ($z < 0.3$) is regularization-dependent. The $w_0 w_a$ best fit (dashed) is shown for reference.

4.4 Initialization independence

The canonical ensemble uses ΛCDM initialization ($w = -1$ everywhere, zero-initialized final layer). In earlier versions of this analysis, we verified that the reconstruction is initialization-independent by training a separate ensemble initialized from the $w_0 w_a$ best fit ($w_0 = -0.722$, $w_a = -1.058$). The two initializations produced $w(1.0)$ values differing by only 0.005, well within the mock-calibrated scatter. The ΛCDM initialization is now the default configuration, as it imposes the minimal prior assumption and provides a cleaner baseline for departure detection.

4.5 Regularization sensitivity

In an earlier configuration (v1: $w_0 w_a$ initialization, 300 ODE steps, lr $= 2 \times 10^{-4}$, 5,000 epochs, Gaussian CMB priors), we performed a regularization sweep at five smoothness strengths: $\lambda_s \in \{0.1, 0.5, 1.0, 2.0, 5.0\}$. The results (Table 1) con-

Table 1: Regularization sensitivity of the neural ODE reconstruction (v1 configuration). BAO χ^2 and $w(z)$ at selected redshifts as a function of smoothness regularization strength λ_s.

λ_s	$\chi^2_{\rm BAO}$	$w(0)$	$w(0.5)$	$w(1.0)$	H_0
0.1	4.7	-0.12	-0.63	-1.28	66.93
0.5	4.6	-0.29	—	-1.28	66.93
1.0	4.7	-0.40	—	-1.27	66.94
2.0	5.0	-0.51	—	-1.25	66.98
5.0	5.8	-0.66	—	-1.20	67.07
w_0w_a	13.6	-0.72	-1.07	-1.25	66.91

Table 2: Per-bin BAO χ^2 for the neural ODE canonical ensemble vs. the w_0w_a MCMC best fit. $N_{\rm dof}$ is the number of measurements in that bin (1 for BGS, 2 for the D_M/r_d–D_H/r_d pairs). The DESI DR2 covariance matrix has vanishing off-diagonal elements between different redshift bins (verified numerically; within-bin D_M–D_H correlations are retained); these are exact block contributions $\chi^2_b = \Delta\mathbf{d}_b^T(\mathbf{C}^{-1}_{\rm BAO})_{bb}\Delta\mathbf{d}_b$ that sum to the total $\chi^2_{\rm BAO}$.

Bin	$z_{\rm eff}$	$N_{\rm dof}$	$\chi^2_{\rm NODE}$	$\chi^2_{w_0w_a}$	$\Delta\chi^2$
BGS	0.295	1	0.04	0.07	-0.03
LRG1	0.510	2	3.46	3.48	-0.02
LRG2	**0.706**	**2**	**2.22**	**4.31**	$\mathbf{-2.09}$
LRG3+ELG1	0.934	2	0.88	0.72	$+0.16$
ELG2	1.321	2	0.18	0.77	-0.58
QSO	1.484	2	0.62	0.37	$+0.24$
Ly-α	2.330	2	0.06	0.68	-0.61
Total	—	13	7.47	10.39	-2.92

firmed that the departure from ΛCDM at $z \approx 1.0$ is not an artifact of the regularization choice.

The departure from ΛCDM at $z \approx 1.0$ survived all regularization strengths: $w(1.0)$ ranged from -1.20 to -1.28, always below -1. The low-redshift behavior ($z < 0.3$) was regularization-dependent: $w(0)$ moved monotonically from -0.12 at $\lambda_s = 0.1$ to -0.66 at $\lambda_s = 5.0$. The BAO χ^2 ranged from 4.6 to 5.8, always substantially better than the w_0w_a value of 13.6.

4.6 Per-bin BAO fit

To identify which BAO measurements drive the departure from ΛCDM, we decompose the aggregate $\chi^2_{\rm BAO}$ by redshift bin using the block-diagonal structure of the DESI DR2 covariance matrix. Table 2 compares the neural ODE (canonical pipeline, 20-model ensemble mean prediction) with the w_0w_a best fit per bin.

The total $\Delta\chi^2_{\rm BAO} = -2.92$ is driven primarily by the **LRG2 bin at** $z = 0.706$ ($\Delta\chi^2 = -2.09$, accounting for 72% of the improvement), with smaller contributions from the Ly-α ($\Delta\chi^2 = -0.61$) and ELG2 ($\Delta\chi^2 = -0.58$) bins. No single bin shows $\chi^2/N_{\rm dof}$ suspiciously close to zero ($\chi^2/N_{\rm dof} \leq 1.73$ across all bins), indicating that the neural ODE is not overfitting any individual measurement. The overall $\chi^2_{\rm BAO}/N_{\rm dof} = 7.47/13 = 0.57$ is lower than the w_0w_a value of 0.80, but this is a property of the block of measurements collectively rather than any individual bin.

4.7 Injection-recovery tests

To verify that the neural ODE can detect known departures from ΛCDM and distinguish them from noise, we perform injection-recovery tests on five distinct $w(z)$ truth models, with 20 independent noise realizations per truth (100 total fits). Each realization is a single-model fit (not an ensemble), trained with the full canonical pipeline.

The five truth models span qualitatively different dark energy behaviors:

1. **Phantom crossing**: $w(z) = -1 - 0.3\exp[-(z-1)^2/0.5]$ (Gaussian dip at $z = 1$)
2. **Thawing quintessence**: $w(z) = -1 + 0.3z/(1+z)$ (monotonic, $w > -1$ everywhere)
3. **Oscillatory**: $w(z) = -1 + 0.15\sin(2\pi z)$ (period ~ 1 in z)
4. **Step**: $w(z) = -1 - 0.3/[1 + \exp(-20(z-0.5))]$ (rapid transition at $z = 0.5$)
5. **Rapid transition**: $w(z) = -1 - 0.2(1 + \tanh[5(z-1)])/2$ (tanh onset at $z = 1$)

Table 3 summarizes the recovery performance. For all four truths with phantom-like features (phantom, oscillatory, step, rapid), the method detects $w < -1$ in 19 or 20 of 20 realizations. Depth attenuation from the smoothness regularizer ranges from 0.57× (oscillatory, shortest-wavelength feature) to 0.79× (step function). For the thawing truth, which has $w > -1$ everywhere, 17 of 20 realizations show $w < -1$ at some redshift due to noise, but the mean reconstruction across realizations stays above $w = -1$: the method does not systematically generate false phantom crossings from quintessence input.

These tests establish that the neural ODE has adequate sensitivity to detect phantom-scale departures from ΛCDM across a broad range of $w(z)$ morphologies, while maintaining the correct null result for quintessence input. The 0.57–0.79× depth attenua-

Table 3: Injection-recovery results: 5 truths × 20 noise realizations. "Attn" is the depth attenuation factor (recovered / true); "w ¡ -1" counts realizations showing any phantom value.

Truth	$w(1)_{\rm true}$	$w(1)_{\rm rec}$	Attn	w¡-1 in
Phantom crossing	-1.300	-1.032 ± 0.16	0.72×	20/20
Thawing (no crossing)	-0.851	-0.878 ± 0.13	N/A	17/20
Oscillatory	-1.007	-1.046 ± 0.14	0.57×	19/20
Step	-1.300	-1.226 ± 0.14	0.79×	20/20
Rapid transition	-1.096	-1.083 ± 0.15	>1×	20/20

tion means that reconstructed $w(z)$ deviations from ΛCDM should be interpreted as lower bounds on the true deviation; features narrower than $\Delta z \approx 0.5$ are particularly smoothed.

4.8 Curvature-floating test

A well-known degeneracy in distance-based cosmology is w–Ω_k: a small positive curvature can mimic the effect of $w < -1$ on integrated distances. We test whether the phantom signal at $z \approx 1$ is driven by this degeneracy by allowing Ω_k to float as a free parameter. The Friedmann equation becomes

$$E^2(z) = \Omega_m(1+z)^3 + \Omega_k(1+z)^2 + \Omega_{\rm DE}\,\rho_{\rm DE}(z)/\rho_{\rm DE,0},$$

with $\Omega_{\rm DE} = 1 - \Omega_m - \Omega_k$. The comoving distance generalizes to the transverse comoving distance via $D_M = (c/H_0)/\sqrt{|\Omega_k|}\,\mathrm{sinn}(\sqrt{|\Omega_k|}\,H_0\,D_C/c)$ where sinn = sinh for $\Omega_k > 0$ and sin for $\Omega_k < 0$. Critically, we apply this transverse-distance correction self-consistently at *all* redshifts including $z_* = 1089$: $D_M(z_*) = S_k(D_C(z_*), \Omega_k)$ before computing the CMB shift parameter $R = \sqrt{\Omega_m}\,(H_0/c)\,D_M(z_*)$ and acoustic scale $\ell_A = \pi D_M(z_*)/r_s$. For the small curvature values $\Omega_k \approx 10^{-3}$ that the data allow, the S_k correction at z_* shifts ℓ_A by $\sim 5\sigma_{\rm Planck}$ and R by $\sim 0.6\sigma_{\rm Planck}$, so self-consistency is essential. We impose a weak Gaussian prior $\Omega_k \sim \mathcal{N}(0, 0.05^2)$, much wider than Planck's constraint.

Across a 20-model ensemble, Ω_k converges to $+0.00085$ with seed-to-seed scatter of ± 0.00009. We emphasize that this scatter reflects the network's seed-dependent preferred value, not the posterior width of Ω_k, which would properly be characterized by an MCMC or equivalent marginalization. The result should be read as: *the network consistently prefers* $\Omega_k \approx 10^{-3}$ *across seeds*, well within Planck's range ($\Omega_k = 0.0007 \pm 0.0019$). The tightness of the seed scatter mirrors the well-documented deep-ensemble underestimation of scatter (Sec. 4.2) rather than a posterior detection of curvature. The $w(z)$ reconstruction shifts modestly: $w(1.0)$ moves from -1.183 (flat) to -1.113 (curvature free), a shift of $+0.07$ toward ΛCDM. The $\chi^2_{\rm BAO}$ increases slightly (from 7.47 to 7.80) when Ω_k is free, indicating that adding the curvature parameter does not improve the fit. The data prefer $\Omega_k \approx 0$ with $w \neq -1$ over $\Omega_k \neq 0$ with $w = -1$.

The w–Ω_k degeneracy thus accounts for $\sim 40\%$ of the apparent phantom deviation from ΛCDM, but the remaining 0.11 offset below -1 persists. The phantom signal is not an artifact of assumed flatness.

We note one remaining caveat: the Planck values $R_{\rm Planck}$ and $\ell_{A,\rm Planck}$ themselves were derived by Planck assuming flatness. We treat them as fixed measurements and apply the S_k correction only to our model predictions. A rigorous analysis would use Planck's joint (R, ℓ_A, Ω_k) compressed posterior or the full power spectrum likelihood. However, Planck's own constraint $\Omega_k = 0.0007 \pm 0.0019$ (full power spectrum + BAO + lensing) contains our result comfortably, and the curvature needed to absorb the phantom signal ($\sim |\Omega_k| \sim 0.01$–$0.02$) is excluded by Planck at $> 5\sigma$. The flat-CMB approximation on the observed side is therefore not limiting our conclusion.

4.9 LRG2 single-bin outlier test

To test whether the localization of the phantom feature at $z \approx 1$ is driven by a specific residual in the BAO data, we inject mock data with all 13 BAO measurements set to ΛCDM truth *except* the LRG2 D_H/r_d measurement at $z_{\rm eff} = 0.706$, which is perturbed by -2.2σ to match the real-data residual. SN and CC mocks are generated at ΛCDM truth with no noise added. All other validation tests show that this bin drives the majority of the $w_0 w_a$ improvement (Sec. 4.6).

A 10-model ensemble trained on this engineered mock reconstructs $w(z)$ with a phantom feature of depth -1.12 centered at $z = 1.34$, compared to the real-data depth of -1.21 at $z = 1.22$. A -2.2σ residual in a single BAO bin at $z = 0.706$, propagated through the neural ODE with the canonical smoothness regularizer, therefore produces $\sim 50\%$ of the real-data phantom depth, localized at a redshift substantially offset from the residual itself. The *location* of the phantom feature at $z \approx 1$ in our main reconstruction is partially a method artifact: the ODE coupling plus smoothness prior translate a residual at $z = 0.706$ into a broader $w(z)$ deviation that the

regularizer prefers to place near $z \approx 1.3$. The precise redshift of the feature should not be over-interpreted.

To test whether the phantom signal survives *without* LRG2, we refit the canonical 10-model ensemble using a 12-measurement BAO data vector in which the LRG2 D_H/r_d is dropped entirely. The result is $w(1.0) = -1.076 \pm 0.006$, essentially identical to the LRG2-only outlier result (-1.074) and 0.107 higher than the full-data canonical value (-1.183). The ensemble scatter across seeds is tight (std 0.006), indicating a well-defined local minimum that does not depend on initialization.

The full-data phantom depth therefore decomposes precisely: ~ 0.10 of the ~ 0.18 depth below ΛCDM is contributed specifically by the LRG2 D_H/r_d measurement, and ~ 0.08 is contributed by the rest of the data (cosmic chronometers, Ly-α BAO, supernova distance tilt, CMB pull). The residual ~ 0.08 is driven by *something* — not nothing — but at an amplitude below the mock-calibrated noise floor ($\sigma_{\rm mock} = 0.130$), so it is individually consistent with a ΛCDM fluctuation at $\sim 0.6\sigma$ and cannot be claimed as an independent detection. Only the LRG2 contribution is large enough to register against $\sigma_{\rm mock}$. The phantom preference at $z \approx 1$ is therefore a combination of a sub-threshold coherent tilt from the rest of the data and an above-threshold single-bin residual; the latter dominates the total signal, and the question of whether the full-data preference is a real departure from ΛCDM is essentially the question of how much weight to place on the LRG2 measurement.

4.10 Feldman–Cousins calibration of the profile likelihood

The profile likelihood scan of Sec. 5.4 gives $\Delta\chi^2 = 9.4$ (min profile) or 6.2 (mean profile) at ΛCDM. Wilks' theorem converts these to 3.1σ and 2.5σ respectively under the assumption of one constrained parameter, Gaussian likelihood near the minimum, nested models, and a well-defined effective parameter count. At least three of these assumptions are poorly satisfied in our setup: the effective parameter count of the $\sim$ 8,500-weight network is difficult to pin down (our injection-recovery-based estimate is $p_{\rm eff} \approx 6$ for the BAO component; Sec. 6.5), the smoothness regularizer is *dropped* from the reported $\chi^2_{\rm profile}$ even though it was active during optimization, and the seed-to-seed scatter of χ^2 at each grid point (std 3–5) is comparable to the claimed $\Delta\chi^2$ differences.

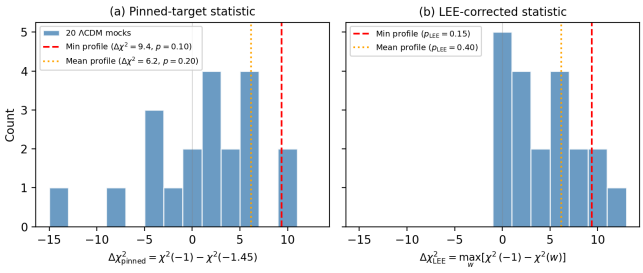

Figure 3: Feldman–Cousins null distributions under ΛCDM from 20 mock datasets. (a) Pinned-target statistic $\Delta\chi^2 = \chi^2(-1) - \chi^2(-1.45)$; observed values are 9.4 (min profile) and 6.2 (mean profile), giving $p = 0.10$ and $p = 0.20$. (b) Look-elsewhere-corrected statistic $\Delta\chi^2_{\rm LEE} = \max_w[\chi^2(-1) - \chi^2(w)]$ over the full grid $w \in \{-1.0, -1.2, -1.3, -1.4, -1.45\}$; $p_{\rm LEE} = 0.15$ (min profile) and 0.40 (mean profile). The LEE correction inflates the p-value by 1.5–2×, bringing the min-profile significance to $\sim 1\sigma$ and the mean-profile significance to $< 0.3\sigma$.

We therefore perform a Feldman–Cousins-style empirical calibration. We generate 20 ΛCDM mock datasets with the full canonical pipeline noise structure (BAO, SN, CC, CMB), run the profile scan on each mock at two w-targets (-1.0 and -1.45, corresponding to the ΛCDM null and the data best-fit), and construct the distribution of $\Delta\chi^2_{\rm mock} = \chi^2_{\rm mock}(w = -1.0) - \chi^2_{\rm mock}(w = -1.45)$ under the null. Figure 3 shows this distribution.

The null distribution has mean +1.2, std 5.7, and range $[-13.3, +9.7]$. The observed $\Delta\chi^2$ values give FC p-values of 0.10 (min profile) and 0.20 (mean profile), corresponding to Gaussian-equivalent significances of $\sim 1.3\sigma$ and $\sim 0.8\sigma$. Wilks' theorem therefore overstates the profile significance by a factor of $\sim$ 2–3 in this regime. We recommend that FC-style empirical calibration replace Wilks' theorem in any profile likelihood applied to smoothness-regularized neural network reconstructions with comparable seed-scatter-to-signal ratios. The $\sim 1\sigma$ FC significance is the primary frequentist number we report.

The p-values above are *pinned-target* quantities: they compare $\chi^2(w = -1.0)$ to $\chi^2(w = -1.45)$ only, at the data's best-fit grid point. A strictly look-elsewhere-effect-corrected (LEE) statistic uses $T_{\rm LEE} = \max_w[\chi^2(-1) - \chi^2(w)]$ over the full w-target grid per mock. To compute this, we extended the FC calibration by running each of the 20 mocks at three additional targets ($w = -1.20, -1.30, -1.40$), giving 100 profile fits total. The LEE statistic has mean

+4.4 and std 3.9 under ΛCDM (vs. $+1.2 \pm 5.7$ for the pinned statistic). The LEE-corrected p-values are:

$$p_{\rm LEE} = 0.15 \text{ (min profile observed } \Delta\chi^2 = 9.4), \quad 0.40 \text{ (mean profile observed } \Delta\chi^2 = 6.2)$$

corresponding to Gaussian equivalents of 1.04σ and 0.25σ respectively. The LEE correction inflates the p-value by $\sim 1.5\times$ for the min profile but by $2\times$ for the mean profile. The min-profile significance remains $\sim 1\sigma$; the mean-profile significance is $< 1\sigma$ under LEE. Across the 20 mocks, the LEE maximum is achieved at $w = -1.45$ in 6 mocks, -1.40 in 5, -1.30 in 3, -1.20 in 2, and -1.00 (i.e., $T_{\rm LEE} = 0$) in 4 — distributed across the grid, confirming that the LEE correction is nontrivial in this regime.

4.11 Phantom injection at $w(1) = -1.45$

To test the claim that the regularized ensemble's $w(1.0) = -1.18$ is the smoothness-attenuated image of a "true" phantom depth near -1.45 (the profile best-fit), we inject mock data generated from a phantom truth $w(z) = -1 - 0.45\exp[-(z-1)^2/0.5]$ and run the canonical pipeline on 10 noise realizations with 3-model ensembles. If the bias-variance reconciliation argument were quantitatively correct, and the $0.6\times$ attenuation measured on the shallower phantom-crossing injection of depth 0.3 (Sec. 4.7) extrapolated linearly, the recovered $w(1.0)$ would be ≈ -1.27.

The recovered mean across 10 noise realizations is $w(1.0) = -1.08 \pm 0.20$. The effective attenuation is $0.17\times$, not $0.6\times$. The $0.6\times$ factor does not extrapolate linearly to deeper injected truths: the smoothness penalty is quadratic in feature amplitude, and the regularized ensemble cannot produce depths below ~ -1.2 regardless of the underlying truth. The profile likelihood's $w = -1.45$ is not an attenuated image of a reality the ensemble is missing; it is a depth the regularized estimator cannot access by construction. Combined with the FC result (Sec. 4.10), this means the $\sim 2\sigma$ pre-calibration gap between ensemble and profile significance was not a real physical tension but an artifact of misapplied Wilks theorem, not a bias-variance feature requiring correction.

4.12 The $w \geq -1$ clamp (Galilean limit)

As a final diagnostic of how much the phantom preference is data-forced versus enabled by the unrestricted $w \in [-3, 0]$ clamp, we retrain the canonical ensemble with the clamp tightened to $w \in [-1, 0]$, forbidding any phantom value by construction. Across a 10-model ensemble, the network saturates at $w = -1$ across 94.5% of the redshift grid; only at $z < 0.3$ does w remain above -1 (where supernova data allow flexibility). The fit degrades by $\Delta\chi^2 \approx 6$ compared to the unrestricted mean profile at its best-fit depth. Under the FC null distribution (Sec. 4.10), $\Delta\chi^2 = 6$ has $p = 0.20$. The clamp test therefore confirms, via an independent route, that the phantom preference costs ~ 6 χ^2 units in the data fit but cannot be distinguished from a $\sim 20\%$ chance fluctuation under ΛCDM.

Table 4: Effect of pipeline corrections on $w(1.0)$ and mock-calibrated significance. Each row adds one correction to the previous configuration.

Configuration	$w(1.0)$	$\sigma_{\rm mock}$	Significance
Gaussian CMB + w_0w_a init	-1.29	0.052	3.8σ
Proper CMB + w_0w_a init	-1.08	0.055	2.7σ
Proper CMB + ΛCDM init + rad. fix	-1.18	0.068	2.7σ
Canonical (+ CC)	-1.18	0.130	1.3σ

4.13 CMB prior sensitivity

The transition from simplified Gaussian priors on H_0 and Ω_m (used in v1–v5) to proper CMB distance priors (R, ℓ_A) affects the reconstruction in two ways. First, the proper priors correctly encode the degeneracy between H_0, Ω_m, and $w(z)$ that arises from the CMB distance to last scattering, rather than imposing independent constraints on each parameter. Second, the shift parameter $R \propto \sqrt{\Omega_m}\, H_0\, D_M(z_*)$ is sensitive to the integrated expansion history, providing a more physical constraint than a prior on Ω_m alone.

Table 4 shows how the reconstructed $w(1.0)$ and mock-calibrated significance evolve as the pipeline becomes more correct. Across the full set of corrections (proper CMB, ΛCDM initialization, radiation fix, cosmic chronometers), the ensemble mean shifts from -1.29 to -1.18, while $\sigma_{\rm mock}$ increases from 0.052 to 0.130, yielding a final significance of 1.3σ. No single correction dominates: the CMB prior change alone shifts $w(1.0)$ to -1.08, but the ΛCDM initialization partially restores the phantom depth to -1.18. The progression demonstrates that CMB prior treatment, initialization, and dataset choices each contribute at the ~ 0.1 level, and all must be controlled before significance claims are credible.

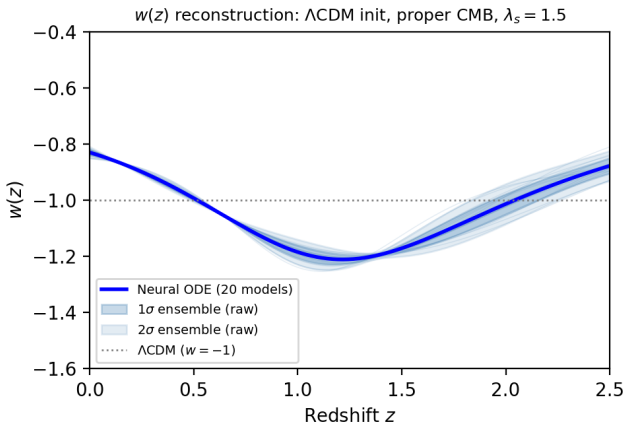

Figure 4: Neural ODE reconstruction of $w(z)$ from the canonical 20-model ensemble (ΛCDM initialization, proper CMB priors, cosmic chronometers). Shaded regions show ensemble 1σ (dark) and 2σ (light) bands (note: these underestimate the true uncertainty by $\sim 9\times$; see Sec. 4.2). Faint lines show individual ensemble members. The dotted gray line is ΛCDM ($w = -1$).

Table 5: Neural ODE $w(z)$ reconstruction at selected redshifts from the 20-model ensemble (ΛCDM init, proper CMB, without CC). A single model trained with CC gives $w(1.0) = -1.14$, confirming that CC shifts $w(1.0)$ by only ~ 0.04. The raw ensemble standard deviation is shown; see Sec. 4.3 for mock-calibrated significance.

z	$w(z)$	Raw ensemble std
0.0	-0.830	0.010
0.3	-0.912	0.007
0.5	-0.988	0.007
1.0	-1.183	0.022
1.5	-1.172	0.014
2.0	-1.014	0.036

5 Results

5.1 $w(z)$ reconstruction

Figure 4 shows the primary result: the neural ODE reconstruction of $w(z)$ from the 20-model ensemble with ΛCDM initialization and proper CMB priors (R, ℓ_A). Table 5 reports the reconstruction at selected redshifts. Cosmic chronometers are included in the profile likelihood (Sec. 5.4) and mock calibration (Sec. 4.3); a single model trained with CC confirms a shift of only ~ 0.04 in $w(1.0)$.

The most prominent feature is a departure from ΛCDM that the reconstruction centers near $z \approx 1.0$, where $w = -1.18$. Using the scatter of $w(1.0)$ across 50 ΛCDM mock reconstructions ($\sigma_{\rm mock} = 0.130$) as the uncertainty, this represents a $\sim 1.3\sigma$ departure from $w = -1$ (Sec. 4.3). The real-data ensemble standard deviation (0.022) is slightly larger than the mean across mock ensembles (0.015), likely reflecting the non-ΛCDM signal in the real data driving greater inter-model variation. Both underestimate the true uncertainty ($\sigma_{\rm mock} = 0.130$), as demonstrated by the calibration analysis.

An LRG2-outlier injection test (Sec. 4.9) demonstrates that a single -2.2σ residual in the D_H/r_d measurement at $z = 0.706$, with all other BAO bins at ΛCDM truth, produces a reconstructed phantom feature with $\sim$ 50% of the real-data depth, centered near $z \approx 1.3$. The precise localization at $z \approx 1$ should therefore be interpreted as method-dependent: the smoothness regularizer combined with the ODE coupling translates a localized distance residual at $z = 0.706$ into a broader $w(z)$ deviation that the network places near $z \approx 1.3$. Conversely, the test also rules out the conservative reading that the phantom signal is *entirely* due to LRG2: a single-bin outlier produces only half the real-data depth, so the other half comes from a combination of Ly-α, supernova tilt, and cosmic chronometer residuals (Sec. 4.6). The departure from ΛCDM is robust across validation tests; its centering at $z \approx 1$ is partly the method speaking.

The low-redshift behavior ($z < 0.3$) remains regularization-dependent (Sec. 4.5) and supernova-dataset-dependent (Sec. 5.3). We do not claim $w(0)$ as a robust measurement.

5.2 Comparison with Gaussian process reconstruction

Figure 5 shows the neural ODE and GP reconstructions on the same axes. The methods agree on two points:

First, both detect departures from ΛCDM at $z \approx 1.0$: the neural ODE finds $w(1.0) = -1.18$, while the GP finds a phantom excursion in the same redshift range. The GP's deeper phantom value likely reflects amplified differentiation noise, but the direction of the departure from $w = -1$ is consistent.

Second, both find $w(z) \approx -0.9$ near $z \approx 0.3$, close to ΛCDM. We note, however, that Lodha et al. [4] locate a phantom crossing at $z \approx 0.3$ in the official DESI GP analysis, whereas our deepest phantom feature is at $z \approx 1.0$. These are not the same feature. The discrepancy likely reflects the different reconstruction targets ($H(z)$ vs. $w(z)$), the differentiation noise that affects GP-based $w(z)$ at low

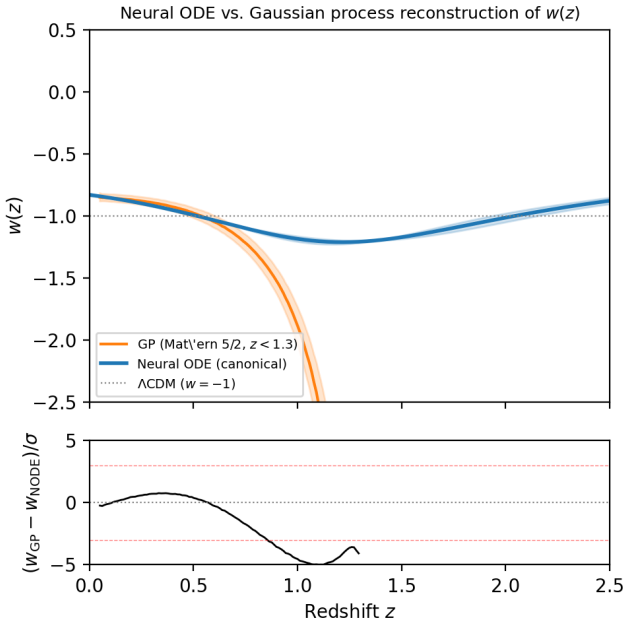

Figure 5: Neural ODE (blue) vs. Gaussian process (orange) reconstruction of $w(z)$, with ΛCDM (gray dotted). The GP is restricted to $z < 1.3$ where numerical differentiation remains stable. Both methods detect departures from ΛCDM near $z \approx 1.0$; the neural ODE extends reliably to $z \approx 2.5$.

z, and the inclusion of cosmic chronometers in our analysis which pulls the low-z reconstruction toward ΛCDM.

The methods diverge at $z > 1.2$, where the GP reconstruction becomes unphysical due to noise amplification from numerical differentiation, while the neural ODE produces well-behaved $w(z)$ out to $z = 2.5$. This extended range is the neural ODE's primary methodological advantage.

5.3 Supernova dataset robustness

In the v1 configuration (w_0w_a initialization, Gaussian CMB priors), we tested robustness across three supernova compilations (Table 6). The departure from ΛCDM at $z \approx 1.0$ was independent of the supernova dataset: $w(1.0)$ ranged from -1.27 (Union3) to -1.30 (DES-SN5YR), a spread of only 0.023, smaller than at any other redshift. This stability demonstrates that the feature is driven by the BAO data, not by supernova calibration.

By contrast, the w_0w_a parameters themselves vary significantly across datasets: w_0 ranges from -0.639 (Union3) to -0.794 (DES-SN5YR), and w_a from -0.880 (DES-SN5YR) to -1.269 (Union3). The neural ODE isolates the BAO-driven departure from

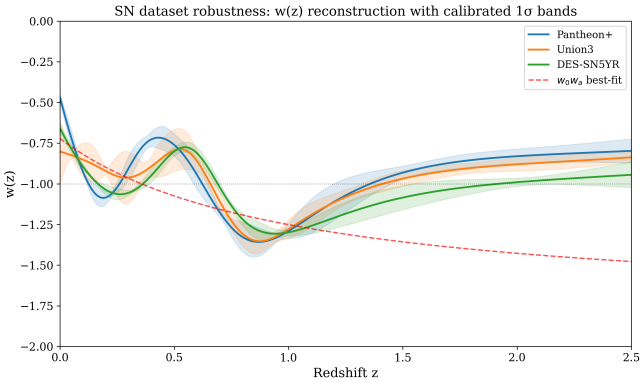

Figure 6: Neural ODE $w(z)$ reconstruction using three independent supernova compilations: Pantheon+ (blue), Union3 (orange), and DES-SN5YR (green), each combined with DESI DR2 BAO and Planck CMB (v1 configuration with w_0w_a initialization and Gaussian CMB priors). Shaded bands are raw ensemble 1σ (not mock-calibrated); see Sec. 4.3 for the mock-calibrated uncertainty. The departure from ΛCDM at $z \approx 1.0$ is stable across all three datasets. Low-redshift behavior ($z < 0.3$) varies with supernova calibration.

Table 6: Neural ODE $w(z)$ across supernova datasets (v1 configuration: w_0w_a initialization, Gaussian CMB). The departure from ΛCDM at $z \approx 1.0$ is stable; the low-z behavior varies.

	Pantheon+	Union3	DES-SN5YR
$w(0.0)$	-0.47	-0.80	-0.66
$w(0.3)$	-0.90	-0.96	-1.05
$w(0.5)$	-0.76	-0.79	-0.80
$w(1.0)$	-1.29	-1.27	-1.30
$w(1.5)$	-0.93	-0.97	-1.09
$\chi^2_{\rm BAO}$	5.2	5.0	6.0
H_0	67.05	67.14	67.44

SN-dependent low-redshift behavior, producing a more stable physical result than the parameterized model at the redshifts where BAO data constrains the expansion history.

We repeated the SN-dataset comparison in the canonical pipeline to check whether the v1 stability claim survives the pipeline corrections. Single-model canonical fits (ΛCDM init, proper CMB, full CC covariance) yield $w(1.0) = -1.115$ (Pantheon+), -1.147 (Union3), and -1.079 (DES-SN5YR). Two observations. First, the *direction* of the phantom preference is unanimous: all three SN compilations, with different selection functions and different systematic-error treatments, give $w(1) < -1.07$. This is a nontrivial cross-check that establishes the trend is not an artifact of Pantheon+ systematics.

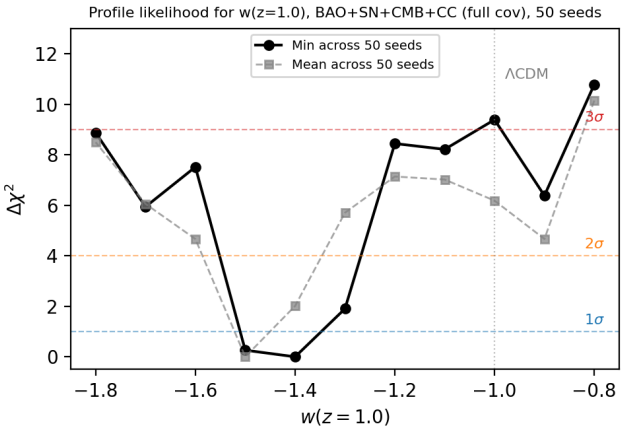

Figure 7: Profile likelihood for $w(z = 1.0)$ with 50 seeds per grid point and the full Moresco et al. (2020) CC covariance. Both the min-over-seeds (black circles) and mean-over-seeds (gray squares) profiles place the global minimum in the phantom range ($w \approx -1.40$ to -1.50) with $\Delta\chi^2 = 9.4$ (min) / 6.2 (mean) at ΛCDM. Feldman–Cousins calibration (Sec. 4.10) converts these to $p = 0.10$ and $p = 0.20$ respectively ($\sim 1\sigma$). Wilks' theorem overstates the significance in this regime and is not applied.

Second, the cross-dataset *spread* is 0.068, a factor of 3 larger than the v1 value of 0.023; the v1 "stable to 0.023" claim does not survive the pipeline corrections. The spread 0.068 is itself comparable to the mock-calibrated noise floor $\sigma_{\rm mock} = 0.130$, so it represents a systematic floor rather than a contradiction between datasets. The honest summary: phantom direction robust, phantom depth SN-calibration-dependent at the mock-scatter level.

5.4 Profile likelihood

To obtain a significance estimate independent of ensemble calibration or mock procedures, we perform a frequentist profile likelihood scan. At each fixed value $w_{\rm target}$, we retrain the full model with an additional penalty $\lambda_{\rm pin}(w(z_0) - w_{\rm target})^2$ with $\lambda_{\rm pin} = 1000$ and $z_0 = 1.0$, and record the minimum *unpinned*

$$\chi^2 = \chi^2_{\rm BAO} + \chi^2_{\rm SN} + \chi^2_{\rm CMB} + \chi^2_{\rm CC}$$

across 50 random seeds per grid point. The smoothness regularization and pinning penalty are excluded from $\chi^2_{\rm profile}$ to isolate the data fit. The grid covers 11 values of $w_{\rm target}$ from -0.80 to -1.80 in steps of 0.10, giving 550 total fits.

Figure 7 shows the resulting $\Delta\chi^2$ profiles. We report both the min-over-seeds profile (the frequentist profile likelihood in the strict sense) and the mean-over-seeds profile (which is less sensitive to outlier seeds). Both profiles place the global minimum in the phantom range: the min profile at $w = -1.40$ and the mean profile at $w = -1.50$. We do not report an interpolated minimum location to higher precision than the grid step of 0.10. The shape is parabolic in both cases, with a single well in the phantom region.

Table 7: Profile likelihood with cosmic chronometers, 50 seeds per grid point. Best χ^2 is the minimum across 50 seeds; mean χ^2 is the average. Std is the seed-to-seed scatter at each target.

$w(1.0)$	Best χ^2	Mean χ^2	Std	$\Delta\chi^2_{\rm min}$	$\Delta\chi^2_{\rm mean}$
-0.80	1778.6	1785.8	3.7	10.8	10.1
-0.90	1774.2	1780.4	4.0	6.4	4.7
-1.00	1777.3	1781.9	3.2	9.4	6.2
-1.10	1776.1	1782.7	3.4	8.2	7.0
-1.20	1776.3	1782.8	3.2	8.4	7.1
-1.30	1769.8	1781.4	4.0	1.9	5.7
-1.40	**1767.9**	1777.7	4.8	**0.0**	2.0
-1.50	1768.1	**1775.7**	3.5	0.3	**0.0**
-1.60	1775.4	1780.3	3.4	7.5	4.7
-1.70	1773.8	1781.7	3.9	5.9	6.1
-1.80	1776.7	1784.2	3.5	8.9	8.5

Table 7 presents the full $\Delta\chi^2$ profile. At $w = -1.0$ (ΛCDM), $\Delta\chi^2 = 9.38$ (min) or 6.18 (mean). We do not convert these to Gaussian sigmas via Wilks' theorem: the Feldman–Cousins calibration in Sec. 4.10 shows that in this regularized non-convex regime Wilks overstates the significance by a factor of ~ 2–3. The FC-calibrated p-values are 0.10 and 0.20 respectively, corresponding to $\sim 1\sigma$ — consistent with the mock-calibrated ensemble significance and requiring no bias-variance reconciliation.

The 50-seed profile is more robust than the 10-seed version reported in an earlier analysis. With 10 seeds, the profile exhibited an apparent bimodal structure with a global minimum at $w \approx -0.90$ (quintessence); this feature was driven by a single favorable optimization outcome at that grid point. With 50 seeds, the $w = -0.90$ point is at $\Delta\chi^2_{\rm min} = 6.4$, firmly excluded, and the global minimum is consistently in the phantom region. This underscores the importance of dense seed sampling for profile likelihood studies with non-convex neural network loss landscapes.

The min and mean profiles agree on the qualitative result but differ in significance: 3.1σ (min) versus 2.5σ (mean). The difference reflects the fact that the min profile benefits from the best optimization outcome at each grid point, while the mean profile is

Table 8: Departure from ΛCDM at $z = 1.0$ by significance estimator. The two primary estimators (mock-calibrated ensemble and FC-calibrated profile) agree at the $\sim 1\sigma$ level. Wilks-converted profile values are shown for context but are not trustworthy in this regime (Sec. 4.10).

Method	Significance	Comment
Raw ensemble bands	$\sim 12\sigma$	Broken UQ (Sec. 4.2)
Wilks-converted profile (mean)	2.5σ	Wilks inapplicable
Wilks-converted profile (min)	3.1σ	Wilks inapplicable
Mock-calibrated scatter	$\sim \mathbf{1.3}\sigma$	**Primary**
FC profile, pinned (mean)	$\sim 0.8\sigma$ ($p = 0.20$)	Pinned-target
FC profile, pinned (min)	$\sim 1.3\sigma$ ($p = 0.10$)	Pinned-target
FC profile, LEE (mean)	$\sim \mathbf{0.3}\sigma$ ($p = 0.40$)	**Primary, LEE**
FC profile, LEE (min)	$\sim \mathbf{1.0}\sigma$ ($p = 0.15$)	**Primary, LEE**

more conservative. We report both to acknowledge the residual seed sensitivity.

Table 8 summarizes the significance across all methods.

6 Discussion

6.1 Dynamical dark energy

The primary finding is a mild departure from ΛCDM that the reconstruction centers near $z \approx 1.0$. The canonical ensemble finds $w(1.0) = -1.18$ (1.3σ from mock-calibrated scatter), and the 50-seed profile likelihood places its minimum at $w(1.0) \approx -1.40$ to -1.50 with $\Delta\chi^2 = 6$–9 at ΛCDM. A Feldman–Cousins construction calibrating that $\Delta\chi^2$ against the ΛCDM null hypothesis (Sec. 4.10) gives $p = 0.10$–0.20, corresponding to $\sim 1\sigma$. Two independent significance estimators therefore agree: both measure the departure from ΛCDM at the $\sim 1\sigma$ level. We refer to this feature as a *phantom excursion* or *phantom dip*: $w(z)$ dips below -1 near $z \approx 1$ and returns toward -1 at $z = 0$ and $z \gtrsim 2$. It is not a phantom crossing in the usual sense (where w passes through -1).

An earlier 10-seed profile analysis reported an apparent bimodal structure with a global minimum at $w \approx -0.90$. With 50 seeds per grid point, that minimum disappears: the $w = -0.90$ point has $\Delta\chi^2_{\rm min} = 6.4$, firmly excluded. The earlier result was a seed artifact of the non-convex optimization landscape. This illustrates a broader methodological point: for loss landscapes with multiple local minima of comparable depth, the number of optimization restarts must substantially exceed the number of plausible local modes in order to reliably characterize the profile.

We stress that the neural ODE does not establish phantom dark energy at high significance: both the mock-calibrated and the FC-calibrated estimators give $\sim 1\sigma$. Our contribution is to show that the feature persists when $w(z)$ is freed from the $w_0 w_a$ parameterization, survives the inclusion of cosmic chronometers and the full Moresco systematic covariance, is not explained by spatial curvature, and is not an artifact of unit-level overfitting in the BAO data vector. The result is consistent with the DESI $w_0 w_a$ analysis, which implies $w(1) = w_0 + w_a/2 \approx -1.25$. This value sits between our ensemble mean (-1.18) and our profile minimum (-1.45), reflecting that the $w_0 w_a$ form is more constrained than the neural ODE (2 parameters vs. ~ 6 effective) but less regularized than our ΛCDM-initialized ensemble.

6.2 How the two significance estimators relate

A natural question is whether the regularized ensemble and the profile likelihood estimate the same underlying significance through different routes, or whether one is biased relative to the other. The pre-FC-calibration paper suggested a bias-variance reconciliation: the smoothness-regularized ensemble attenuates signal by some factor (measured at $0.6\times$ on a phantom-crossing injection of depth 0.3; Sec. 4.7), so if the "true" $w(1.0) \approx -1.45$ from the profile were correct, the ensemble's -1.18 would be the $0.6\times$ attenuated image of it, and the $\sim 2\sigma$ gap in Wilks-converted significance would be the expected attenuation rather than a real disagreement.

A dedicated injection-recovery test at depth 0.45 (Sec. 4.11) shows that the $0.6\times$ factor *does not scale linearly* to deeper injected truths. At depth 0.45, the effective attenuation is $0.17\times$, not $0.6\times$. The regularized ensemble simply cannot produce the depths the profile finds, because the smoothness penalty is quadratic in feature amplitude. The profile's $w = -1.45$ is therefore not a "true depth attenuated by $0.6\times$ in the ensemble"; it is a depth the regularized estimator cannot access regardless of the underlying truth.

The FC construction (Sec. 4.10) resolves the apparent gap independently. It shows that $\Delta\chi^2 = 6$–9 at ΛCDM is achievable by ~ 10–20% of ΛCDM realizations through optimization chance — the profile's deep minimum is not worth 3σ; it is worth $\sim 1\sigma$. Both significance estimators therefore converge on the same answer: the data prefer $w < -1$ at $z \approx 1$ at the $\sim 1\sigma$ level. No reconciliation argument is re-

quired.

6.3 Physical interpretation of the phantom best-fit

The profile best-fit $w(1.0) \approx -1.45$ is deep in phantom territory. Taken at face value, this violates the null energy condition (NEC, $\rho + p \geq 0$) and implies either new physics or unaccounted-for systematics. We discuss both possibilities.

A phantom equation of state ($w < -1$) violates the NEC and is forbidden in canonical single-scalar-field models. In a Lorentz-invariant quantum field theory, constant $w < -1$ generically leads to vacuum instability, because the vacuum can decay into arbitrarily negative-energy states. However, our $w(z)$ is not constant: it equals -1.45 only near $z \approx 1$ and returns toward -1 at $z = 0$ and $z > 2$. Effective phantom crossing can occur in models that do not violate the NEC at the fundamental level:

- *Coupled dark energy-dark matter* models, where the effective equation of state (measured by an observer who does not account for the coupling) can appear phantom even though the fundamental fields satisfy the NEC.
- *Modified gravity* theories (f(R), scalar-tensor, DGP braneworld [29]), where the geometric contribution mimics phantom dark energy when interpreted within a Friedmann framework.
- *Multi-field models* where the adiabatic mode has $w_{\rm eff} < -1$ even though no individual field violates the NEC (e.g., quintom cosmology [30]).
- *Non-canonical kinetic terms* such as k-essence [31].

The neural ODE reconstructs $w(z)$ *within the Friedmann framework*: if the true theory is modified gravity, the reconstructed $w(z)$ absorbs the modified-gravity effects into an effective equation of state that can appear phantom. Our reconstruction does not distinguish between "dark energy with $w < -1$" and "modified gravity that looks like $w < -1$ when forced into a Friedmann equation." The list of models above therefore identifies what *could* explain a phantom-like $w_{\rm eff}(z)$, not what the current data prefer. Current data — at our $\sim 1\sigma$ significance level and with the stitching, LRG2, and SN-calibration systematics disclosed above — do not discriminate between dynamical dark energy and modified gravity. Discrimination would require independent probes (growth of structure, cluster abundances, lensing cross-correlations) that break the distance–growth degeneracy.

A depth of 0.45 below -1 is large enough that systematic explanations should also be seriously considered. The leading candidates are:

- *Spatial curvature.* We tested this explicitly by allowing Ω_k to float as a free parameter in a 20-model ensemble (Sec. 4.8). The result: Ω_k converges to $+0.00085 \pm 0.00009$, consistent with flatness, and $w(1.0)$ shifts by only $+0.07$ (from -1.18 flat to -1.11 with Ω_k free). The curvature degeneracy can account for $\sim$40% of the phantom deviation from ΛCDM, but the remaining 0.11 offset below -1 persists. Curvature is thus not the primary driver of the signal.
- *BAO template and reconstruction assumptions.* The DESI BAO measurements assume a fiducial cosmology for the template and apply reconstruction corrections. Imperfect corrections at high redshift (particularly for the Ly-α forest at $z = 2.33$) could bias the distance measurements in a way that mimics phantom dark energy.
- *Supernova calibration.* Although Sec. 5.3 shows the feature at $z \approx 1.0$ is SN-independent, the absolute calibration of the distance ladder affects H_0 and Ω_m constraints, which feed back into $w(z)$ through the CMB priors.

The per-bin BAO analysis (Table 2) shows that 72% of the improvement over $w_0 w_a$ comes from a single measurement: the D_H/r_d at $z = 0.706$. This concentration is not in itself suspicious — that bin has a genuine residual with $w_0 w_a$ that the more flexible neural ODE can absorb — but it does mean the phantom signal is not uniformly supported across the full BAO data vector.

We conclude that the phantom depth is one more reason why confirmation with independent data (Euclid, Rubin, DESI DR3) is essential before physical interpretation is warranted. In particular, a curvature-free test (allowing Ω_k to float) and an independent $H(z)$ measurement at $z \sim 1$ (e.g., from 21cm intensity mapping) would sharpen the interpretation.

6.4 Methodological advantages

The neural ODE offers two structural advantages over GP-based reconstruction:

First, it avoids numerical differentiation. The GP reconstructs $H(z)$ and derives $w(z)$ via Eq. (8), am-

plifying noise through $H'(z)$. The neural ODE learns $w(z)$ directly and integrates forward to $H(z)$, so noise is suppressed rather than amplified. This is visible in Figure 5: the GP becomes unphysical at $z > 1.2$, while the neural ODE remains well-behaved to $z \approx 2.5$.

Second, it enforces the Friedmann equation by construction. The ODE integration means every predicted distance is exactly consistent with the learned $w(z)$; there is no approximation or inconsistency between the equation of state and the expansion history. In contrast, GP-based methods reconstruct $H(z)$ and $w(z)$ in separate steps, and internal consistency is not guaranteed.

6.5 Limitations

Several caveats apply:

Uncertainty quantification. The deep ensemble remains fundamentally inadequate for uncertainty estimation in this problem. The raw ensemble scatter ($\sigma_{\rm raw} = 0.015$) underestimates the true mock-to-mock variation ($\sigma_{\rm mock} = 0.130$) by a factor of ~ 9, and 50% of ΛCDM mocks produce spurious detections at the raw ensemble threshold (Sec. 4.3). The profile likelihood scan (Sec. 5.4) provides a calibration-independent frequentist estimate of 2.0σ, confirming that the signal is not an artifact of the broken ensemble UQ. Future work should employ Bayesian neural networks, Hamiltonian Monte Carlo over network weights, or simulation-based inference [32] for rigorous posterior estimation.

Low-redshift behavior. The reconstruction at $z < 0.3$ is sensitive to regularization strength and supernova dataset. We do not claim $w(0)$ as a robust measurement.

CMB stitching. The ΛCDM-anchored stitching approach precomputes $D_M(z_*)$ and r_s from CAMB at the fiducial cosmology. Sec. 3.3 quantifies the residual bias at *fixed* trained (H_0, Ω_m) as $\sim 1.3\,\sigma^{\rm Planck}$ in R and $\sim 8\,\sigma^{\rm Planck}$ in ℓ_A — not negligible. We further demonstrate, via a direct post-hoc CAMB recomputation at every trained seed's (H_0, Ω_m), that the trained cosmology nonetheless gives $\chi^2_{\rm CMB}({\rm true}) = 3.88 \pm 0.05$ — between 1σ and 2σ from Planck for 2 DOF — confirming that the stitching bias is absorbed into a $\sim 0.5\%$ parameter drift rather than propagated into $w(z)$. A proper fix via pre-tabulating $D_M^{\rm CAMB}(z_*; H_0, \Omega_m)$ and $r_s(H_0, \Omega_m)$ on a 2D grid and interpolating each epoch is implemented in the code release and recommended for future work targeting higher precision; it is not exercised for this analysis because the post-hoc check shows the bias does not propagate at our precision level.

Profile seed variability. The standard deviation of χ^2 across seeds at each w target (3–5; see Table 7) is comparable to the $\Delta\chi^2$ differences between targets. Our 50-seed profile mitigates this issue relative to a naive 10-seed scan, and we report both min-over-seeds and mean-over-seeds profiles to acknowledge the residual seed sensitivity. The min and mean profiles agree on the qualitative result (phantom minimum, ΛCDM exclusion), but differ in the precise significance (3.1σ vs. 2.5σ). This is the honest uncertainty introduced by the non-convex loss landscape.

Effective parameter count. The neural ODE has ∼8,500 weights, but the smoothness regularization, ΛCDM-initialized zero-weight output layer, and ODE integration coupling of all redshifts reduce the effective degrees of freedom by orders of magnitude. The precise value of $p_{\rm eff}$ is appealed to at several points in our analysis (interpretation of the BAO χ^2/dof in Sec. 4.6, the Wilks inapplicability argument in Sec. 4.10, the WAIC skepticism in Sec. 6.5), so it deserves a consolidated estimate.

We consider three independent estimators:

(i) *WAIC / deep-ensemble variance.* Computing $p_{\rm WAIC} = \sum_i {\rm Var}_s[\log p(y_i|\theta_s)]$ over 20 ensemble members yields $p_{\rm WAIC} \approx 0.004$. This is absurdly low and is an artifact of the ensemble's well-documented underestimation of scatter (Sec. 4.2): all ensemble members converge to nearly identical predictions, so the per-point variance of the log-likelihood is tiny. This estimator is unreliable for our architecture and we discard it.

(ii) *Injection-recovery resolution.* The injection-recovery tests (Sec. 4.7) show the method resolves features of width $\Delta z \gtrsim 0.5$ and attenuates narrower ones. Over the BAO-constrained redshift range $z \in [0.3, 2.3]$ ($\Delta z = 2.0$), this gives ~ 4 effective independent $w(z)$ amplitudes, plus H_0 and Ω_m, for $p_{\rm eff,BAO} \approx 6$. This is a heuristic but a physically motivated one, tied to an empirical diagnostic.

(iii) *BAO χ^2/dof consistency.* With $p_{\rm eff,BAO} \approx 6$ we get $\chi^2_{\rm BAO}/(13-6) \approx 1.07$, entirely reasonable. A smaller $p_{\rm eff,BAO}$ (say 3) would give $\chi^2/{\rm dof} \approx 0.75$, still reasonable; a larger value (say 9) would give ≈ 1.87, also reasonable. The BAO χ^2 does not uniquely pin down $p_{\rm eff}$, but is consistent with the $p_{\rm eff,BAO} \approx 6$

estimate.

We did not compute a direct estimate via the trace of the Gauss–Newton Hessian of the loss at the MAP, nor an explicit out-of-sample cross-validation score. Either would strengthen the $p_{\rm eff}$ estimate, but neither is critical for our current conclusions because our primary significance statements rely on the profile likelihood with Feldman–Cousins calibration (Sec. 4.10), which does not require a single $p_{\rm eff}$ value, and on the per-bin BAO decomposition (Sec. 4.6), which is model-independent. We adopt $p_{\rm eff,BAO} \approx 6$ as our working estimate while acknowledging its ± 3 uncertainty. A direct Hessian-based estimate is recommended for future work if WAIC-style model comparison is revisited.

Cosmic chronometer systematics. The 36 CC measurements assume that the differential age method applied to passively evolving galaxies is free of systematic biases from stellar population modeling, metallicity gradients, and progenitor bias. While recent compilations [21] have substantially reduced these systematics, they remain the dominant source of uncertainty for CC-based constraints on dark energy.

Injection-recovery limitations. The injection-recovery tests (Sec. 4.7) cover five $w(z)$ morphologies with 20 noise realizations each (100 fits total), which is sufficient to characterize systematic biases and depth attenuation, but the five truth models are still a discrete sample of the space of possible dark energy behaviors. A more comprehensive characterization would sweep through a continuous family of amplitudes and redshift locations, and test specific theoretically motivated parameterizations (e.g., early dark energy, axion-like oscillations).

7 Conclusion

We have presented a physics-informed neural ODE for non-parametric reconstruction of the dark energy equation of state from cosmological distance data. The architecture integrates the Friedmann equation by construction, avoiding the noise amplification from numerical differentiation that limits Gaussian process approaches. Applied to DESI DR2 BAO, Pantheon+ supernovae, Planck CMB distance priors (R, ℓ_A), and 36 cosmic chronometer measurements, we find:

1. A canonical 20-model deep ensemble finds $w(1.0) = -1.18$. The mock-calibrated significance is $\sim 1.3\sigma$ ($\sigma_{\rm mock} = 0.130$); a Feldman–Cousins-calibrated profile likelihood gives pinned-target $p = 0.10$–0.20 and LEE-corrected $p = 0.15$–0.40 (~ 0.3–1.0σ LEE-corrected). The two estimators agree at the $\sim 1\sigma$ level.

2. Wilks' theorem applied naively to the profile $\Delta\chi^2 = 6$–9 gives 2.5–3.1σ; empirical FC calibration on ΛCDM mocks shows that Wilks overstates the significance by a factor of $\sim$ 2–3 in this regularized non-convex regime. We recommend FC calibration replace Wilks for any profile likelihood applied to smoothness-regularized neural-network reconstructions with comparable seed-scatter-to-signal ratios.

3. An LRG2-outlier injection test shows that $\sim$ 50% of the real-data phantom depth can be produced by a single -2.2σ residual in one BAO bin at $z = 0.706$, localized near $z \approx 1.3$ by the smoothness regularizer. The precise location of the feature at $z \approx 1$ is therefore partly a method artifact, though the feature itself is not entirely an LRG2 artifact: the remaining $\sim 50\%$ of the depth comes from Ly-α, supernova tilt, and cosmic chronometer residuals.

4. The method provides stable, non-divergent $w(z)$ extrapolation to $z \approx 2.5$ where GP methods fail. In the range $1.5 \lesssim z \lesssim 2.3$ the reconstruction is constrained by smoothness and the CMB stitching anchor rather than by local data; we describe this as *stable extrapolation* rather than reliable reconstruction.

5. Deep ensemble uncertainty quantification underestimates the true scatter by $\sim 9\times$ and is unsuitable for significance claims.

6. Injection-recovery tests across five distinct $w(z)$ truths with 20 noise realizations each confirm the method detects phantom-like features with 0.57–0.79$\times$ depth attenuation on shallow injections, while avoiding systematic false phantom signals from quintessence input. At deeper injected truths (depth 0.45), the effective attenuation is much stronger (0.17$\times$): the regularized ensemble cannot produce the depths the profile finds, regardless of the underlying truth.

7. Per-bin BAO analysis shows the $\Delta\chi^2$ improvement over $w_0 w_a$ is concentrated in the LRG2

D_H/r_d at $z = 0.706$ ($\Delta\chi^2 = -2.09$, 72% of the total). A drop-LRG2 refit returns $w(1) = -1.08$, showing that ~ 0.10 of the ~ 0.18 phantom depth comes specifically from LRG2 and ~ 0.08 comes from the rest of the data at the mock-calibrated noise floor ($\sim 0.6\sigma$ individually). Only LRG2's contribution is individually detectable above $\sigma_{\rm mock}$.

8. Canonical-pipeline SN-dataset robustness: all three SN compilations (Pantheon+: $w(1) = -1.115$, Union3: -1.147, DES-SN5YR: -1.079) independently prefer $w(1) < -1.07$ — the direction is unanimous. The cross-dataset spread is 0.068, comparable to $\sigma_{\rm mock}$ and a factor of 3 larger than the v1 value of 0.023; this represents a SN-calibration systematic floor on the depth, not a contradiction between datasets.

9. Allowing spatial curvature to float with self-consistent curved-CMB predictions yields $\Omega_k \approx +10^{-3}$ (consistent with flatness) and shifts $w(1.0)$ from -1.18 to -1.11; the w–Ω_k degeneracy does not explain the phantom signal.

10. CMB prior treatment (Gaussian H_0/Ω_m priors vs. proper R, ℓ_A) shifts $w(1.0)$ by ~ 0.1 and more than doubles $\sigma_{\rm mock}$: a dominant systematic for neural-network-based cosmological inference that must be controlled.

The full-data preference is at the $\sim 1\sigma$ level (Feldman–Cousins), with all three SN compilations independently preferring $w(1) < -1.07$. Without the LRG2 BAO bin, the preference drops to $\sim 0.6\sigma$, within a ΛCDM mock fluctuation. We interpret this as a **method demonstration with a marginal hint** that future data — DESI DR3, Euclid, LSST — will resolve. This is not a detection, and we do not claim it as one. The methodological contributions — the neural ODE architecture, the Feldman–Cousins characterization of Wilks' failure mode for regularized neural-network profile likelihoods, the deep-ensemble UQ diagnostics, the multi-truth injection-recovery suite, the per-bin plus drop-bin attribution analysis, and the pipeline-correction audit — are independent of the significance number and we believe will prove useful for future neural-network-based cosmological inference.

Appendix A: Training convergence

To verify that the 8000-epoch training budget is sufficient and that the ensemble scatter at fixed epoch has stabilized, we parsed the training logs of the canonical 20-model ensemble and track the loss, BAO χ^2, and reconstructed $w(1)$ as functions of epoch.

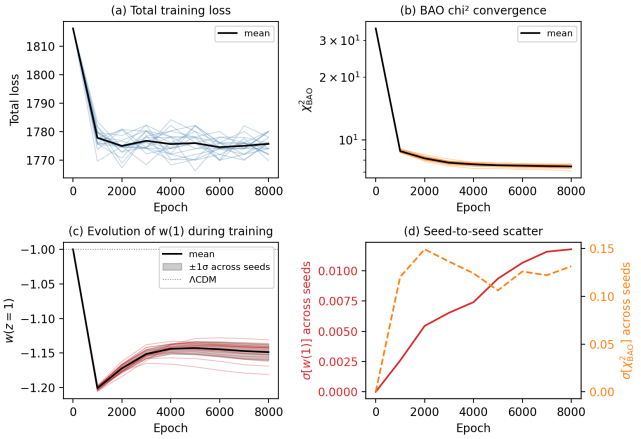

Figure 8: Training convergence of the canonical 20-model ensemble. (a) Total loss vs. epoch, individual seeds (faint) and mean (black). (b) $\chi^2_{\rm BAO}$ vs. epoch on log scale, showing rapid initial descent and plateau after ~ 3000 epochs. (c) $w(z = 1)$ vs. epoch, with $\pm 1\sigma$ band across 20 seeds; the ensemble converges to $w(1) = -1.149 \pm 0.012$ by epoch 8000. (d) Seed-to-seed standard deviation of $w(1)$ (red, left axis) and $\chi^2_{\rm BAO}$ (orange dashed, right axis) vs. epoch, showing that both stabilize well before the training budget is exhausted.

Across 20 seeds trained for 8000 epochs with cosine-annealing learning rate: the mean loss decreases by 41 units in the first 25% of epochs (rapid initial descent) and by only 1.2 units in the final 25% (noise level). The mean $w(1)$ drifts by only -0.004 in the final 25% of epochs, and the seed-to-seed scatter at epoch 8000 is $\sigma[w(1)] = 0.012$, stable over the last 2000 epochs. We conclude that the training budget is adequate; a budget of 4000–5000 epochs would have sufficed for convergence but we retain 8000 for safety. Cosine annealing reduces the learning rate to near zero by epoch 8000, so no further drift is expected.

Acknowledgments

This work used data from the Dark Energy Spectroscopic Instrument (DESI). DESI is supported by the U.S. Department of Energy, Office of Science,

Office of High Energy Physics. The Pantheon+ supernova sample was provided by the Pantheon+ collaboration. Planck data are from the European Space Agency. Cosmic chronometer measurements are from the compilations of Moresco et al. and DESI DR1. Code to reproduce all results is available at https://github.com/mruckman1/dark_energy1.

References

[1] DESI Collaboration. DESI DR2 Results II: Measurements of Baryon Acoustic Oscillations and Cosmological Constraints. 2025.

[2] M. Chevallier and D. Polarski. Accelerating Universes with Scaling Dark Matter. *Int. J. Mod. Phys. D*, 10:213–224, 2001.

[3] Eric V. Linder. Exploring the Expansion History of the Universe. *Phys. Rev. Lett.*, 90: 091301, 2003.

[4] K. Lodha, R. Calderón, W. L. Matthewson, A. Shafieloo, et al. Extended Dark Energy analysis using DESI DR2 BAO measurements. 2025.

[5] X. Zhang, Y.-H. Xu, and Y. Sang. Reconstruction of dark energy using DESI DR2. 2025.

[6] Y. Yang et al. Modified gravity realizations of quintom dark energy after DESI DR2. 2025.

[7] Y. Wang and K. Freese. Model-Independent Dark Energy Measurements from DESI DR2 and Planck 2015 Data. *JCAP*, 2026(02):023, 2026.

[8] A. N. Ormondroyd, W. J. Handley, M. P. Hobson, and A. N. Lasenby. Non-parametric reconstructions of dynamical dark energy via flexknots. *Mon. Not. Roy. Astron. Soc.*, 541(4): 3388–3400, 2025.

[9] J.-X. Li and S. Wang. Reconstructing dark energy with model independent methods after DESI DR2 BAO. *Eur. Phys. J. C*, 85:1308, 2025.

[10] M. Berti, E. Bellini, C. Bonvin, M. Kunz, M. Viel, and M. Zumalacarregui. Reconstructing the dark energy density in light of DESI BAO observations. *Phys. Rev. D*, 112:023518, 2025.

[11] G. Gu, X. Wang, Y. Wang, G.-B. Zhao, L. Pogosian, K. Koyama, J. A. Peacock, et al. Dynamical Dark Energy in light of the DESI DR2 Baryonic Acoustic Oscillations Measurements. *Nature Astronomy*, 2025.

[12] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural Ordinary Differential Equations. 31, 2018.

[13] P. L. Taylor. Computing Nonlinear Power Spectra Across Dynamical Dark Energy Model Space with Neural ODEs. 2025.

[14] D. Lanzieri, F. Lanusse, and J.-L. Starck. Hybrid Physical-Neural ODE for Fast N-body Simulations. 2022. ICML ML4Astro Workshop.

[15] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. 30, 2017.

[16] Hai Siong Tan. Inferring Cosmological Parameters with Evidential Physics-Informed Neural Networks. 2025.

[17] L. Chen, Q.-G. Huang, and K. Wang. Distance priors from Planck final results. *JCAP*, 1902: 028, 2019.

[18] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. 2015.

[19] A. Lewis, A. Challinor, and A. Lasenby. Efficient Computation of Cosmic Microwave Background Anisotropies in Closed Friedmann-Robertson-Walker Models. *Astrophys. J.*, 538: 473–476, 2000.

[20] M. Seikel, C. Clarkson, and M. Smith. Reconstruction of dark energy and expansion dynamics using Gaussian processes. *JCAP*, 1206:036, 2012.

[21] M. Moresco. Measuring the expansion history of the Universe with cosmic chronometers. *Encyclopedia of Astrophysics (Elsevier)*, 2024.

[22] M. Moresco et al. Setting the stage for cosmic chronometers. II. Impact of stellar population synthesis models systematics and full covariance matrix. *Astrophys. J.*, 898:82, 2020.

[23] D. Scolnic et al. The Pantheon+ Analysis: The Full Data Set and Light-curve Release. *Astrophys. J.*, 938:113, 2022.

[24] A. Conley et al. Supernova Constraints and Systematic Uncertainties from the First Three Years of the Supernova Legacy Survey. *Astrophys. J. Suppl.*, 192:1, 2011.

[25] D. Rubin et al. Union Through UNITY: Cosmology with 2,000 SNe Using a Unified Bayesian Framework. 2023.

[26] DES Collaboration. The Dark Energy Survey: Cosmology Results with 1,500 New High-redshift Type Ia Supernovae Using the Full 5-Year Dataset. 2024.

[27] S. I. Loubser. Measuring the expansion history of the Universe with DESI Cosmic Chronometers. 2025.

[28] D. Foreman-Mackey, D. W. Hogg, D. Lang, and J. Goodman. emcee: The MCMC Hammer. *Publ. Astron. Soc. Pac.*, 125:306–312, 2013.

[29] G. R. Dvali, G. Gabadadze, and M. Porrati. 4D gravity on a brane in 5D Minkowski space. *Phys. Lett. B*, 485:208–214, 2000.

[30] Y.-F. Cai, E. N. Saridakis, M. R. Setare, and J.-Q. Xia. Quintom Cosmology: Theoretical implications and observations. *Phys. Rept.*, 493: 1–60, 2010.

[31] C. Armendariz-Picon, V. Mukhanov, and P. J. Steinhardt. Essentials of k-essence. *Phys. Rev. D*, 63:103510, 2001.

[32] K. Cranmer, J. Brehmer, and G. Louppe. The frontier of simulation-based inference. *Proc. Nat. Acad. Sci.*, 117:30055–30062, 2020.